LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-228502

FY07 I/O Integration Blueprint

D. P. Wiltzius, M. R. Gary

March 1, 2007

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Lawrence Livermore National Laboratory

Integrated Computing and Communications Department

FY07 I/O Integration Blueprint

Reviewed Release v1.13

UCRL-TR-228502

January 2, 2007

K.Cupps	BG/L
K.Fitzgerald	Lustre deployment
P.Hamilton	Purple
B.Loewe	Scalable IO, File Systems
J.Long	Green Data Oasis
M.Seager	Sequoia RFP (Petaflop computer)
J.Shoopman & M. Gary	Archival Storage
J.Slavec	Facility Networks
D.Smith & D.Fox	Network Attached Storage
D.Southard	Visualization
D.Wiltzius	Facility Networks, Interconnects, Special Projects

Other contributors:
John Allen, Tom Connell, Doug East, Brent Gorda, Terry Heidelberg, Bryan Lawver, Steve Louis, Terri Quinn

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Purpose

This document provides an understanding of the near and long term computing and I/O resources in the Secure Computing Facility (SCF) and Open Computing Facility (OCF). Requirements for data flows, storage capacities and transfer rates are determined. Recommendations are made for architectures, timeframes for major deliverables, and procurements for the next fiscal year.

Scope

This document will provide an understanding of the resources creating most of the data (e.g., platforms, compute servers, visualization servers, etc) in the timeframe of next year and up to 5 years out. *The goal will be to recommend architectures, timeframes for major deliverables, and procurements for this fiscal year for OCF and SCF.* All resources that have significant I/O flows for local and tri-Lab remote computing resources and user communities will be considered: platforms, compute servers, visualization servers, archival storage system, network attached storage, site-wide global file systems, and networks. The general outline of this document is as follows:

1) Plans for next year and discussion of 2-5 year plans regarding major drivers of I/O requirements:
   a) platforms (e.g., Peloton deployments, Sequoia preparation and deployment)
   b) compute servers
   c) visualization servers
   d) archives
   e) special strategies, mandates, considerations, projects (e.g., Green Data Oasis (GDO))
2) Current architecture discussion that includes network topology, network attached storage (e.g., NFS servers), archival storage (e.g., HPSS), visualization servers, site-wide global file system (e.g., Lustre), and WAN connectivity to tri-Lab remote resources, collaborators and users. In this section information will also be provided, to the extent possible, on the past and current IO capacity and bandwidth.
3) IO requirements and analysis. This section will also try to summarize and focus the future architecture, capacities and bandwidths, and the steps to achieve the future state.
4) Identification of I/O issues and recommendations for addressing issues to achieve the architecture, capacities, and transfer rates suggested throughout the document.
5) Plans for FY07 and beyond. This will focus on providing major FY07 deliverables, and procurement plans. Longer term milestones and such may also be re-stated or summarized in this section.

Appendices will be used to provide additional information that is relevant to the IO architecture of our HPC facilities, but not necessarily of primary consideration.

- A: Inventory of major resources in the OCF and SCF. Mostly tables and statistics, no analysis.
- B: SCF and OCF computing model. This is our generalized HPC "usage model."

Executive Summary

The most valuable aspect of the FY07 I/O Blueprint effort is the exchange of information, and the integration discussions leading to requirements, design and procurement decisions. The essence of the resultant recommendations and decisions is captured in this document, primarily in sections 4 and 5. Section 4 identifies high level issues and makes recommendations on how to react to or mitigate those issues. This section first addresses strategic issues, then goes on to the tactical issues. Section 5 summarizes large procurements.

To establish a framework for the I/O Blueprint effort, Section 1 summarized the most significant drivers for I/O requirements in FY07, and beyond. Section 2 provides a review of the architecture, which in general evolves slowly over the years. Section 3 is an attempt to identify I/O requirements based on a combination of historical trends and through application of the I/O requirement drivers summarized in Section 1. These sections are supplemented by Appendices that provide additional detail or support information for parts of the document.

For FY07 the most influential factor is the exponential growth of computational resources with Purple, BGL and the Peloton systems going into production, and the anticipated addition of capacity computing clusters and finally the Sequoia RFP. Even though the Sequoia RFP has been delayed, we expect to see a 10 Petaflop computer sited in the LC SCF in 2011. A smaller machine, but still at least 0.5 Petaflops, is expected to arrive in early 2008. As a result of these I/O drivers and other issues identified in the Blueprint, the following key efforts, issues and plans are identified within:

- The Data Storage Group articulated their strategy (Section 4) to meet the bandwidth and capacity requirements for the HPSS archival systems in this era of a Petaflop ecosystem. This strategy requires that new platform purchases either include high-bandwidth front-end nodes or include SLIC (Storage/Lustre Interface Cluster) hardware. Sustained, adequate funding and manpower is required to ensure the successful execution of this strategy to meet the rapidly growing bandwidth and capacity archival storage for these computing resources.

- The Lustre file system is carefully realizing the strategy to provide only a few file systems (ideally two) for each facility. The expected benefit is reduced hardware costs over time, but a risk is that the availability and reliability of the file system over time may be compromised. A shortage of resources to test and develop Lustre is described. Significantly, a planned review of competing file systems is scheduled for early CY07. Further, recent hardware failures yet to be fully analyzed raises concerns regarding the priority given to additional Lustre software development efforts and staffing to improve the operational support of Lustre.

- Networking for the anticipated Lustre file system growth required by the computational resources is exceeding the already rapid growth curve in the networking industry (which has exceeded Moore’s Law for decades). The storage controllers and computers can now handily source and sink one or more 10GigE links. This is forcing the network to a 10GigE edge. Large 10GigE switches are not even on most vendors’ roadmaps, the prices of 10GigE switch ports is dropping slowly and the 10GigE PHY module costs – with the exception of the short range CX4 copper PHY – is holding firm. Startups are trying to meet this challenge, but those solutions are rife with risks and surprises.

- Visualization has two significant challenges. One is providing file systems with adequate space and purge policies to allow our user community to do visualization post-processing without constant transfers to and from the archives. The second visualization challenge is image delivery to the desktop. Currently this is being provided by an esoteric technical solution, RGB fiber optic extenders. A proven alternative solution is desktop image delivery via the network. This solution is wrought with security issues.

While this document describes I/O focused requirements, issues, options, plans and deliverables, its true value is as a catalyst for communication and planning between archival storage, network, platform, Network Attached Storage (NAS), Site Wide Global File Systems (SWGFS), visualization and the I/O Testbed personnel.

1) Major I/O requirement drivers

Major computing and file system resource additions drive I/O requirements for FY07. They include:

1) Deployment of Peloton-based machines: Yana, Zeus and Atlas (OCF); Hopi, Minos, and Rhea (SCF) and associated network and Lustre infrastructure. In FY07 these machines will be upgraded with processors running at 2x the performance. Anticipated but not yet acquired funding may add Peloton based resources to the SCF.

2) Advanced Architecture, Sequoia RFP goal is ~0.5 Petaflop (PF) machine goes out for competitive bid by end of FY08. At the end of CY08 a prototype will be sited. *Because new capacity machines are sited in OCF for up to a year before swinging to SCF, both unclassified and classified center infrastructures must be capable of supporting the new machines.*

3) Green Data Oasis. This is an M&IC funded proposal driven by several programs to serve data to an international user community from the Green network (unclassified and unrestricted; outside the LLNL perimeter firewall for the restricted network, but likely will continue to require firewalls). For example, the Earth System Grid climate simulation output is a growing data set that will approach 100TB in a few years. This data set is to be shared by a large international community. Internet access with a large bandwidth capacity shared by many users is desirable.

4) The DisCom WAN contract has been re-competed. The new contract awarded in mid-FY06 provided 10Gbps from LLNL to SNL/NM, and an additional link at 10Gbps from LLNL to LANL. This is a 4x increase in link bandwidth, and an extra link (LLNL-LANL) for additional fault tolerancy compared to the previous DisCom WAN contract. Due to the lack of 10GigE IP encryptors, 1GigE IP encryptors will be used. The initial deployment in early FY06 of 2x1GigE will be increased in early FY07 to 4x1GigE.

Deployment Timelines

The timelines in **Figure 1** show the deployment schedule for the Peloton machines on OCF and SCF, other machines, and the proposed Sequoia. Milestones for OCF are above the timeline and milestones for SCF are below the timeline.

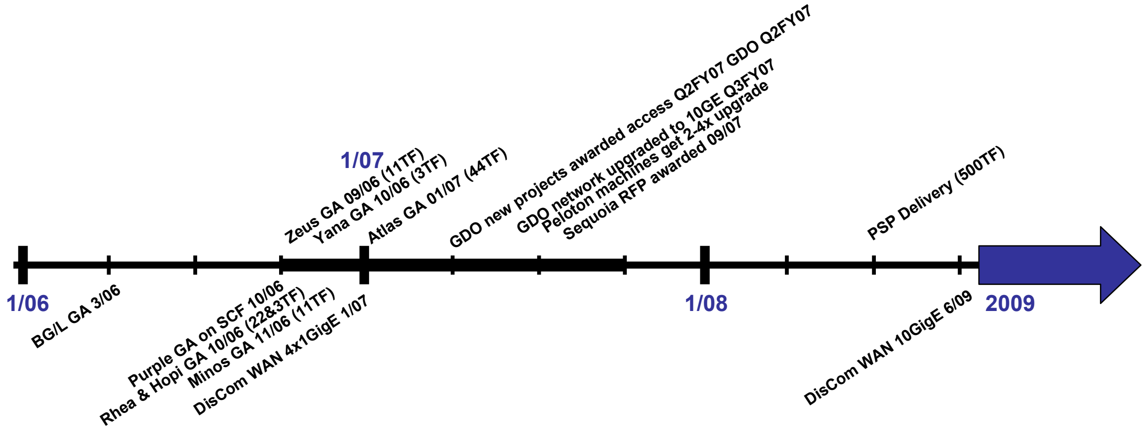


Figure 1: Estimated deployment timelines as of Q1FY07

2) Architecture

The long-term plan for the core I/O Infrastructure is an integrated high performance parallel global file system (i.e., Site Wide Global File System (SWGFS)) and data storage architecture that provides users with fast and uniform access from all the compute resources (see ***Figure 2***). The SWGFS is currently transitioning from a set of local Lustre file systems to a strategy of two site-wide Lustre file systems per facility through the coalescing of Lustre file systems.

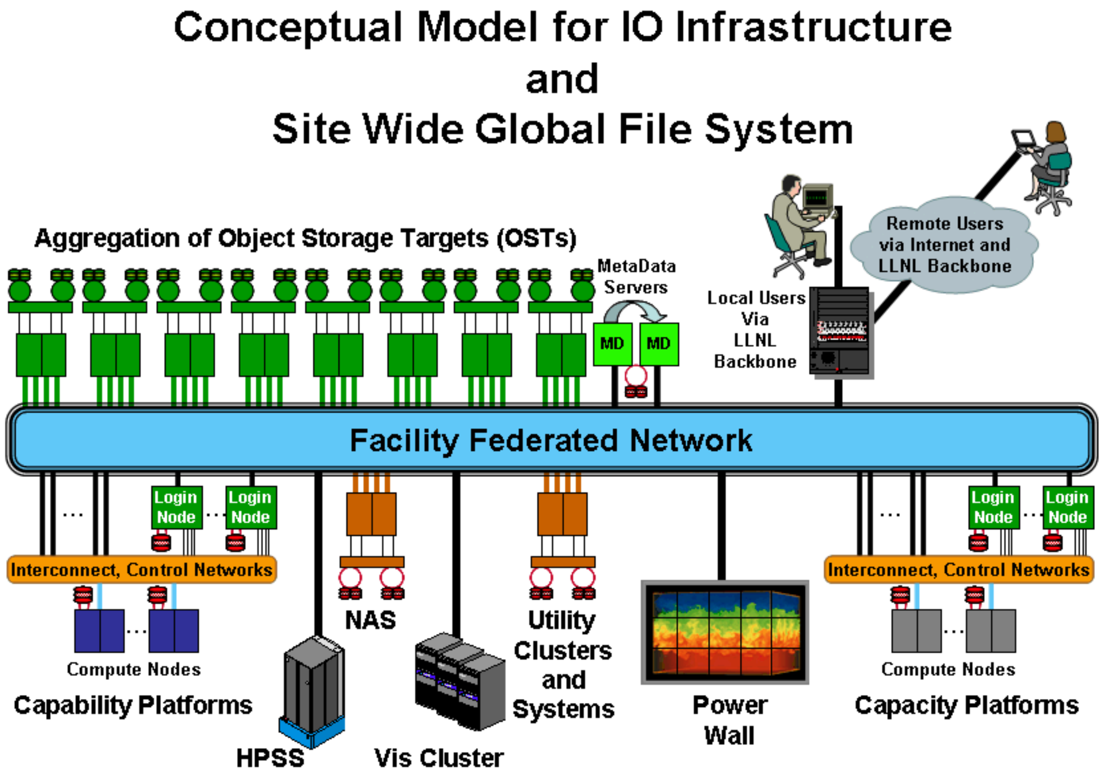

Conceptual Model for IO Infrastructure
and
Site Wide Global File System
Aggregation of Object Storage Targets (OSTs)
MetaData Servers
MD
MD
Local Users Via LLNL Backbone
Remote Users via Internet and LLNL Backbone
Facility Federated Network
Login Node
Login Node
Interconnect, Control Networks
Compute Nodes
Capability Platforms
HPSS
NAS
Vis Cluster
Utility Clusters and Systems
Power Wall
Login Node
Login Node
Interconnect, Control Networks
Compute Nodes
Capacity Platforms

Figure 2: Goal for OCF and SCF facility-wide I/O architecture

The architectures of the platform interconnect and the facility federated network infrastructures can utilize a range of different technologies depending on the preferred approach based on performance, cost and availability. In FY06 LC changed strategy to choose InfiniBand instead of Quadrics as the platform interconnect, 1 and 10 Gigabit Ethernet for the network infrastructure, and Fiber Channel and Serial ATA as the Storage Area Network.

In FY07, the Peloton Linux cluster architecture uses InfiniBand as the interconnect and 10GigE for the LAN and the federated switch Lustre network. Furthermore, rather than converging on one Facility Federated Network we are continuing to deploy federated switch networks for each large platform to provide for the platform high throughput for Lustre, and connectivity to archive, visualization, etc.

The IO architecture for the Open Computing Facility (OCF) and Secure Computing Facility (SCF) is nearly identical (see ***Figure 3***), with only minor differences (e.g., WAN connectivity to Internet (OCF) versus SecureNet (SCF), etc). Both facilities must be capable of supporting each new capability machine as each spends many months in OCF before transitioning to SCF. The Facility Federated Network referred to in ***Figure 2*** is presently used to connect machines with the local Lustre file systems, presently via gateway nodes (not shown) on each machine. In FY07, as per the cost-saving strategy endorsed by the LC Architect, the local Lustre systems will be gradually transitioned to an inter-cluster Lustre file system by cross-mounting the Lustre file systems and coalescing the file system storage and metadata servers.

The sections below will provide an overview of the each subsystem of the extended OCF/SCF facilities. This overview will describe the current architecture, quantify the elements of the sub-system, provide an estimate of the aggregate source/sink throughput rates and capacity, raise issues if appropriate, and provide a list of recommended activities (e.g., performance tests, protocol analysis, etc) that would improve the understanding of the sub-system.

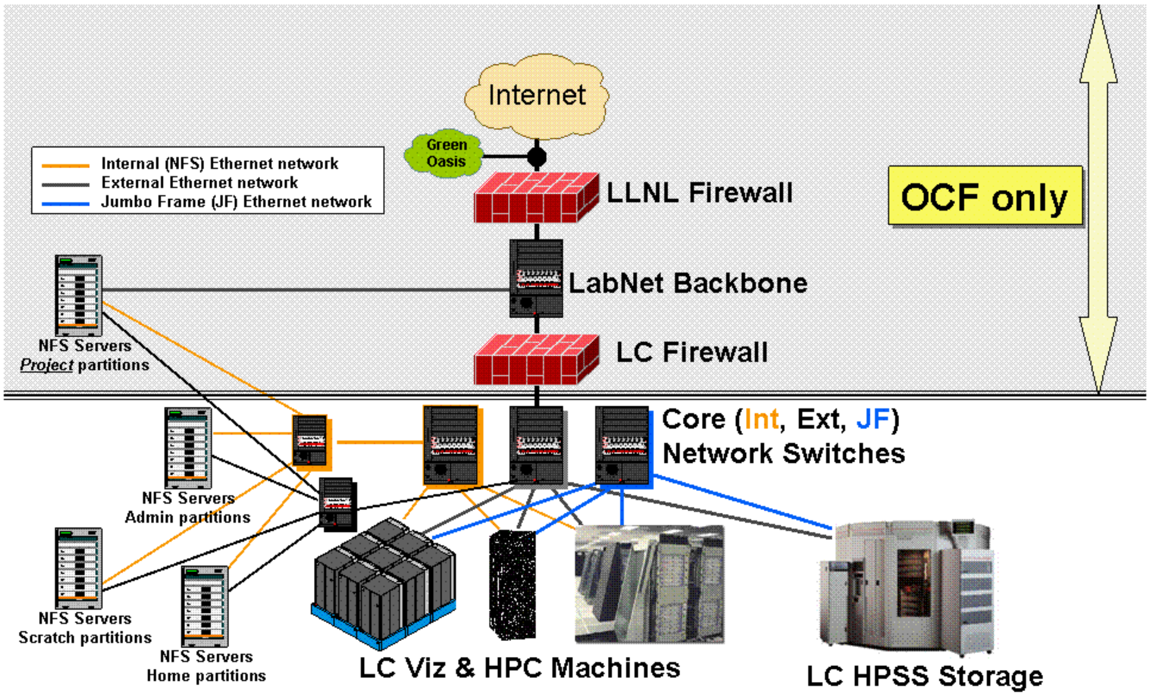


Figure 3: IO architecture overview for FY07

Facility Network

Presently there are three networks in production on the OCF and SCF facilities, as shown in ***Figure 3***: the Internal (or NFS) network for NFS traffic, the External network for inter-facility/desktop traffic, and the Jumbo Frame[1] (JF) network for intra- and tri-Lab facility traffic. In FY07, with consideration of security, total cost of ownership and operational support is the possibility of merging the JF and Internal network. Machines or resources on these networks will have one or more network interface cards (NICs) for that network. For example a typical machine login node will have at least 3 NICs, one each for the Internal, External and JF network. The larger machine login nodes have 2, 4 or even 8 JF NICs for reasons explained below. A fourth network is being deployed: the Facility Federated network for Lustre (hence also referred to as the Lustre network). This network is also configured to support jumbo frames but is designed to meet the high throughput requirements of the platform for the Lustre parallel file system.

Although the OCF and SCF have three or even four networks, these networks do share switches when advantageous (primarily a co-location consideration) to improve switch utilization and costs. Think of this as mostly one physical network with multiple logical networks; that is, virtualization of the networks. Future network products (e.g., upgrades in the IOS) will make network virtualization much easier and simpler. Regardless, the physical separation between OCF and SCF is always maintained, of course.

[1] Jumbo Frame is a non-standard but widely supported feature of Ethernet where the maximum Ethernet packet payload is increased from 1500 to 9000 bytes. Larger packets decrease the load on the host CPU through fewer interrupts and less protocol processing, hence raising the network throughput (significantly (2-3x) with leading edge networks like 10Gigabit Ethernet in FY05) while lowering CPU utilization compared to standard frames.

External network

The External network provides access from outside the OCF/SCF facility for interactive traffic, some file transfers, etc. Specifically, it provides access to OCF/SCF resources for remote and local users via Open/Closed LabNet and the Internet/SecureNet.

Internal network

The Internal network, used for NFS traffic, does not route addresses to or from outside the OCF/SCF facility (e.g., to desktops or LabNet or Internet/SecureNet). This deters many security threats from the broader user community, such as denial of service and breaking into systems.

Jumbo Frame network

The Jumbo Frame network provides high throughput access between tri-Lab facility resources (e.g., HPSS, visualization, HPC machines, clusters) primarily for file transfers, typically utilizing multiple, parallel streams (TCP sockets) across the 4 JF subnets and if necessary over the DisCom WAN to other tri-Lab facilities. “Jumbo Frame” is a poor name for this network since there are other networks that are configured to support jumbo frame Ethernet packets; perhaps a more appropriate name would be “Data Movement Network” or such. The parallelism of data movement is enhanced by machine nodes that have 1 or more NICs on each of the 4 JF subnets. A machine node (e.g., Purple) with multiple NICs on the same JF subnet will be required to bundle these NICs to one logical channel via the IEEE Link Aggregation Control Protocol (802.3ad). On rare occasions, a JF NIC will be configured with an IP address for 2 or 4 of the JF subnets (i.e., aliases) to establish connectivity to each JF subnet without the cost of additional NICs, but this practice is not recommended due to the resultant performance limitation.

	OCF			SCF		
Year	2004	2005	2006	2004	2005	2006
Small core chassis	6	10	9	3	3	3
Large core chassis	17	20	26	11	35	40
1GigE ports	1518	2140	3300	732	3432	3480
10GigE ports	16	40	140	20	98	252

Table 1: Growth of facility core network components and ports in production

The growth of the network for OCF and SCF, both in the number of chassis and number of 1 and 10 Gigabit Ethernet ports, is shown in ***Table 1***. The Facility Federated (Lustre) switch began deployment in the OCF in FY05, and landed in the SCF in mid-FY06; the counts are now merged in the SCF total. Additionally in FY06, consolidation of resources to fewer buildings occurred. These changes impacted the counts between FY05 and FY06.

System and NIC Type	B/w (Gbps) per NIC
Ext	0.25
Int (e.g., NFS client)	0.60
1 GigE JF	0.60
10 GigE JF	5.00
Lustre Gwy/Node 1GigE	0.60
Lustre Gwy/Node 10GigE	5.00
1 GigE OSS/OST	0.50
10 GigE OSS/OST	5.00
NFS Server - NetApp	0.40
NFS Server - BlueArc	0.40
NFS Server - Panasas	0.40

Table 2: Per NIC throughput by network and server type

Source/sink throughput estimates are provided throughout this document based on the per NIC throughput in ***Table 2***. These numbers are estimated simplistically on the basis of the throughput of a NIC for each type of network (e.g., external, internal, jumbo frame, etc) and system (e.g., machine, Lustre client, Lustre OSS/OST, NFS server). This is certainly overly-simplistic until the total network throughput capacity of the client/server is factored (e.g., bandwidth per NIC of a machine with one NIC will likely be more than if the machine had 4 NICs). The throughput of transferring data to/from the machine and media should also be factored in.

DisCom WAN

The DisCom WAN completed a competitive RFP for the OC-48 (2.5Gbps) wide-area network link between LLNL and SNL/NM. With this RFP the Labs moved from the current ATM Fastlane encryptors to HAIPE (High Assurance Internet Protocol Encryption) compliant 1Gbps Taclane (NSA Type 1 IP) encryptors, available since mid-CY2005. 10Gbps Taclanes are not expected to be available before CY2009 (due to lack of demand), so the DisCom WAN will rely on 1Gbps Taclanes for several more years. An IP based WAN is desirable since there are few ATM network components available (we have only one vendor for OC-48 speeds even), ATM product development is stagnant (little R&D funding for these products), and the existing products are challenging to configure to meet our performance requirements. With IP encryptors, we can use mainstream network components for the DisCom WAN.

The new RFP resulted in a dual path 10Gbps WAN with a link from LLNL to Albuquerque via the west coast ("south route", as the previous WAN did) and a second, new link from LLNL to Albuquerque or Santa Fe via Denver ("north route"). LANL has negotiated multiple 10Gbps paths ("Geomax") between LANL and Albuquerque to meet their institutional and ASC (Advanced Simulation and Computing) network requirements. This increase in bandwidth will provide adequate access to the Purple and BG/L machines at LLNL, and when combined with Geomax provides a route diverse, redundant path to LLNL for both Sandia/NM and LANL.

In FY06, LLNL deployed the network monitoring tool Statseeker on the unclassified LC networks, including the DisCom WAN. This tool provides extremely useful information on the utilization of various links. In FY07, security plans are being revised so Statseeker is being deployed on the classified environment. This will provide very useful information on the utilization of the DisCom WAN and the rest of the SCF network.

Network Attached Storage (NAS e.g., NFS)

The Network Attached Storage (NAS) service is from a Gigabit Ethernet NIC connected to the Internal (aka private) network. Typically each login node of each machine has one Internal Gigabit Ethernet NIC. All NFS servers are also connected to the Internal network with one or more Gigabit Ethernet connections (typically 1 or 2, but one NFS server has 60 Gigabit Ethernet ports – see Appendix A for details). The NFS servers only use the NFSv3 protocol, all are configured for NFS over UDP or TCP, and none are configured for jumbo frame.

Most NFS systems also have network connections to the External network. This connection will provide NFS services to machines outside the facility, such as the office machines for LC staff.

On the OCF, some NFS servers are also dual-homed to provide NFS service to the OCF resources and to certain projects. Dual-homed NFS servers have an Internal network connection (for NFS on OCF) and, to provide access for the select projects, an External network connection that goes directly to the LabNet backbone since the LC firewall does not allow NFS traffic to pass through (see ***Figure 3***). This allows projects to share data with their local machines and OCF resources without having to transfer large files (i.e., datasets) in their entirety.

Each NFS server is configured with up to four partitions: Admin, Home, Project, and Scratch. These categories have different performance, reliability, availability and feature set requirements.

Admin Space

The Admin partition is allocated from the NFS servers to LC platforms for system management, configuration management, and shared binaries. An example is the /usr/local file system where a cluster has shared applications, libraries and tools. The system space is fully backed up and has snapshots enabled for easy recovery of files.

Home Space

The Home partition is for user home directories (most users have a quota of 16GB), hence has the highest reliability requirement and so the servers are configured as a fail-over cluster. The files in the home directories tend to be small, so the operations-per-second that a server sustains is the primary performance measure.

The home directory space is backed up to tape monthly, and incremental backups are made daily. Full backups are required for disaster recovery, but the restore time defines how long users are without access to their files. This restore time dictates how large a home directory file system is created. The home directory space has a snapshot or point-in-time-copy feature that lets the user retrieve a file from the file system on demand without needing to go to tape. This is very convenient for both the user and the system administration staff. The snapshot schedule is set at twice a day; noon and 7pm, a total of four snapshots are kept on line. The snapshot feature adds five percent overhead to the usable space in a file system.

Project Space

The Project partition is co-funded by projects for their use with LC computing resources and, as explained above, can be configured for access from the user's local computing environment.

Scratch Space

The Scratch partition is a temporary file storage area for users which is not backed up and is subject to purging. Machines have very little dedicated disk space, so the working space for application results and checkpoints is provided via the NFS scratch partition. Users have also found the center wide scratch space to be a convenient way to share files between capacity and capability clusters.

The primary feature of the scratch space is throughput. The I/O access to this space needs to be fast so that user calculations can continue. The total capacity available is important since the space is purged when the scratch space fills up. To prevent a run-away process from filling up this shared resource, a 100GB quota is imposed on every user. The scratch space is not backed up and snapshots are not enabled. It is the responsibility of the users to save to the archive any data in the scratch space that is needed long term.

NAS capacities

	OCF (TB)			SCF (TB)		
	FY04	FY05	FY06	FY04	FY05	FY06
Admin	4.5	7.3	4.9	2.3	4.0	4.0
Home	18.2	18.2	18.2	10.3	10.3	10.3
Project	13.0	13.8	13.8	7.0	7.0	7.0
Scratch	80.0	93.4	89.3	80.0	96.4	96.4

Table 3 shows the approximate NAS capacity for OCF and SCF for FY04 through FY06. This capacity, and the aggregate bandwidth discussed later, is provided by 18 NFS mounted server systems on OCF and 12 on SCF. On both the OCF and SCF, the largest NFS server is the Panasas NFS server (approx 70TB) and that is for scratch space. The rest of the capacity is provided mostly by Netapp servers but with a few (1-2) Bluearc servers**.**

	OCF (TB)			SCF (TB)		
	FY04	FY05	FY06	FY04	FY05	FY06
Admin	4.5	7.3	4.9	2.3	4.0	4.0
Home	18.2	18.2	18.2	10.3	10.3	10.3
Project	13.0	13.8	13.8	7.0	7.0	7.0
Scratch	80.0	93.4	89.3	80.0	96.4	96.4

Table 3: NFS storage capacity on OCF and SCF for each of the 4 partitions

NFS over UDP vs TCP; jumbo frame

Historically NFS has been based on UDP/IP. One problem with UDP/IP is the poor throughput experienced by NFS clients in environments that drop packets. Dropped packets are particularly disastrous for NFS since NFS generally uses an 8KB datagram, which are then fragmented within six UDP/IP packets that each fit in a (standard frame) 1500 byte Ethernet packet. If one of the Ethernet packets is dropped, the rest are discarded and need to be re-transmitted. Another problem is that jumbo frame Ethernet can only be deployed with UDP/IP in the most carefully controlled network environment, which in practice prevents the use of jumbo frame and the typically significant improvement in throughput achieved with NFS over jumbo frame.

With NFS running over TCP/IP, jumbo frames can then be enabled on the NFS servers. Machines can be configured for jumbo frame or not and the MTU (i.e., packet size) discovery that is part of establishing a TCP connection will determine the proper MTU (e.g., standard or jumbo frame). With jumbo frame, an 8KB NFS datagram will now fit in one jumbo frame Ethernet packet avoiding the inefficiencies experienced with dropped packets. Furthermore, if the NFS over TCP/IP connection does not use jumbo frame the NFS datagram is packaged in several Ethernet packets as with the UDP/IP scenario. However, TCP will re-transmit only the dropped data and not the entire NFS datagram as in the case with NFS over UDP/IP. In conclusion, if the NAS team observes a large number of dropped packets or re-transmitted NFS messages, they should consider running NFS over TCP instead of UDP.

Scalable I/O

The scalable I/O project (SIOP) provides high-performance parallel file system and I/O library support for all major platforms at LLNL. It performs acceptance testing as part of the procurement process, and collaborates with platform partners, academic researchers, and vendors to address ASC high-performance I/O needs.

The SIOP will accept, benchmark, and support LLNL parallel file systems, developing any testing tools (IOR and mdtest) necessary to ensure a stable and high-performance file system. The project provides customer support for parallel file systems, middleware (MPI-IO) and higher-level I/O libraries (HDF5, PnetCDF) and any application I/O issues. With an aim at anticipating future I/O issues, the SIOP will continue to pursue future technologies for file systems with its ASC Alliance contract with UCSC for metadata performance scaling and IETF Open group designs for pNFSv4 and POSIX Extensions.

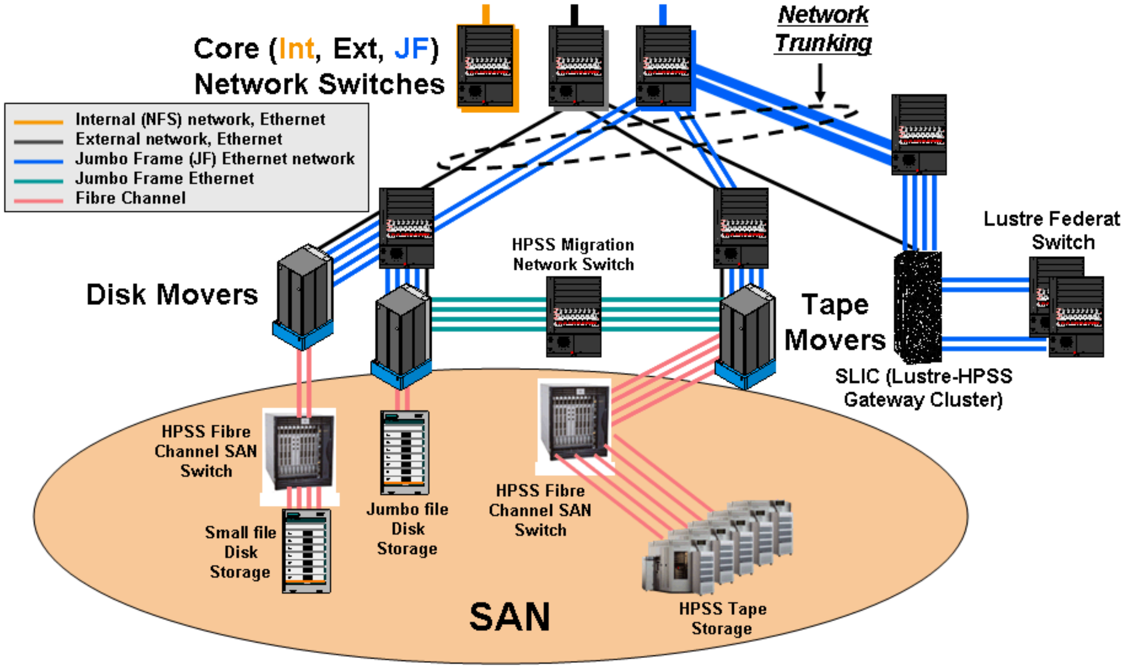

Core (Int, Ext, JF)
Network Switches
Network Trunking
Internal (NFS) network, Ethernet
External network, Ethernet
Jumbo Frame (JF) Ethernet network
Jumbo Frame Ethernet
Fibre Channel
Lustre Federated Switch
HPSS Migration Network Switch
Disk Movers
Tape Movers
SLIC (Lustre-HPSS Gateway Cluster)
HPSS Fibre Channel SAN Switch
Jumbo file Disk Storage
HPSS Fibre Channel SAN Switch
Small file Disk Storage
SAN
HPSS Tape Storage

Figure 4: LC HPSS Storage Sub-system Architecture

Storage (e.g., HPSS)

The HPSS storage system and architecture is largely an independent sub-system of the facility-wide architecture (see ***Figure 3***), requiring adequate network bandwidth for the Internal, External, and Jumbo networks. Furthermore, the HPSS storage architecture (***Figure 4***) is very similar for both OCF and SCF. In this figure, the three core networks (Internal – orange, External – black, and Jumbo Frame - blue) are depicted. Also shown is the HPSS "migration" jumbo frame network used strictly for the HPSS disk and tape "background" migration activity. Finally, the Storage Area Network (SAN) is shown with connections to the IO devices and HPSS movers. The SAN consists of about 60 Brocade Fibre Channel (FC) switches connected to HPSS movers and IO devices by one or more FC Host Bus Adapters (HBA) each with a bandwidth of 2 Gbps (e.g., 250MB/s). Most of the 60 Brocade FC switches are small switches deployed to connect 4-5 tape drives to a mover machine; the rest are larger switches used to share disk controllers with multiple mover machines for transfers of smaller files.

HPSS User Applications

Users read, write and manage their files on HPSS through three user applications: PNFT, HTAR, and PFTP (the Hopper interface runs atop these interfaces). HPSS keeps statistics on each of these three tools. PFTP (Parallel FTP) and FTP (serial FTP) exist on LC machines as one client. The selection of serial or parallel, and the selection of network interfaces on the source and sink machines is determined dynamically and dependent on the source and sink machine and considering the network architecture between those two machines.

PFTP and HTAR were written to use parallel streams load balanced using multiple Jumbo Frame NICs to achieve high throughput for data transfers. PNFT was implemented using PFTP as the underlying data movement mechanism, so PNFT also achieves high throughput by parallel streams over multiple network paths when appropriate. On OCF in FY2006 PFTP was used to move 78% of the files and 25% of the data, PNFT for 21% of the files and 15% of the data, and HTAR for 1% of the files and 60% of the data - but be aware that HTAR stats count as one file an operation on a collection of files (e.g., HPSS jargon is “member files”), so that distribution statistic is inherently skewed against HTAR. On SCF in FY2006 PFTP was used to move 34% of the files and 14% of the data, PNFT for 65% of the files and 40% of the data, and HTAR for 1% of the files and 46% of the data (these are approximate to give an indication of the trend). These data clearly indicate that HTAR continues to move the majority of data on both OCF (60%) and SCF (46%) as compared to other HPSS user applications. Although it is interesting to note that SCF users have found PNFT to be the interface of choice this year in terms of number of files moved, and SCF PNFT users moved nearly as much data (749 TB) as HTAR users (830 TB) . Other tools such as Hopper use one or more of these HPSS user applications to do the data movement, however HPSS does not keep statistics on the use of higher level data management tools such as Hopper.

Parallel Streams and HPSS Movers

As mentioned above data movement applications use multiple, parallel streams over different physical networks for greater throughput. HPSS also employs striping across multiple physical media to achieve higher aggregate throughput. The HPSS movers must coordinate the multiple data streams effectively not only for higher throughput, but also considering availability of data if one or more mover machines become unavailable. For this reason HPSS does striping to multiple devices. In the case of disk transfers, the striped devices for one application are all connected to one mover machine. Hence multiple data streams over the network are intelligently distributed on a mover machine across multiple mover processes, drives and controllers to achieve effective end-to-end parallelization.

Storage Lustre Interface Cluster

On SCF, HPSS also provides a gateway, the Storage Lustre Interface Cluster (SLIC), between HPSS and the Lustre federated network. Users log into the SLIC cluster to FTP files between HPSS and the mounted Lustre file systems. SLIC is a Linux cluster since Lustre is only directly accessible from Linux at this time. Software development efforts are underway in FY05 to provide access to Lustre for HPSS movers for some transfers. Each of the 10 SLIC nodes has four Gigabit Ethernet interfaces for each of the Jumbo Frame and Lustre Federated switch network, and one External Gigabit Ethernet interface – a total of 40 Jumbo Frame Gigabit Ethernet NICs, 40 Gigabit Ethernet NICs for the Lustre federated switch, and 10 Gigabit Ethernet NICs for the External network. The 40 Jumbo Frame NICs on SLIC are connected to an aggregation switch in B453, and trunked to the core Jumbo Frame switch with two 10GigE links. FY06 measurements with Thunder (performed when SLIC was on OCF; Thunder is essentially the same hardware and OS as SLIC) and a high-end dual Xeon Intel processor get over 900Mbps for standard frame Ethernet. Hence the SLIC gateway is capable of substantial network throughput.

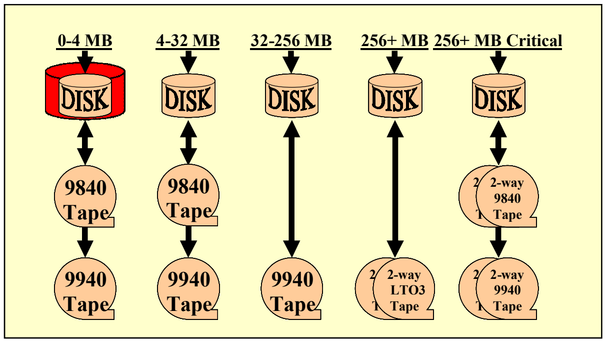


Figure 5: HPSS Storage Hierarchies by Class of Service (COS)

Class of Service (COS) and Storage Hierarchy

HPSS defines 5 classes of service based on the file size: 0-4MB, 4-32MB, 32-356MB, >256MB (256+MB), and 256+MB Critical. ***Figure 5*** illustrates the migration strategy of HPSS based on the COS. For example, the smallest files are migrated from disk to two types of tape, and the largest files are migrated and mirrored to one or, for critical COS files, two types of tape.

Keeping in mind that the y-axis in ***Figure 6*** is logarithmic, notice that the number of files archived in FY06 decreases as the COS file size increases, but the amount archived increases – ignoring the "critical COS." This trend is due to the fact that the HTAR tool concatenates a large number of files (e.g., a sub-directory tree) into one file, and the statistics for writes count this as one (large) file.

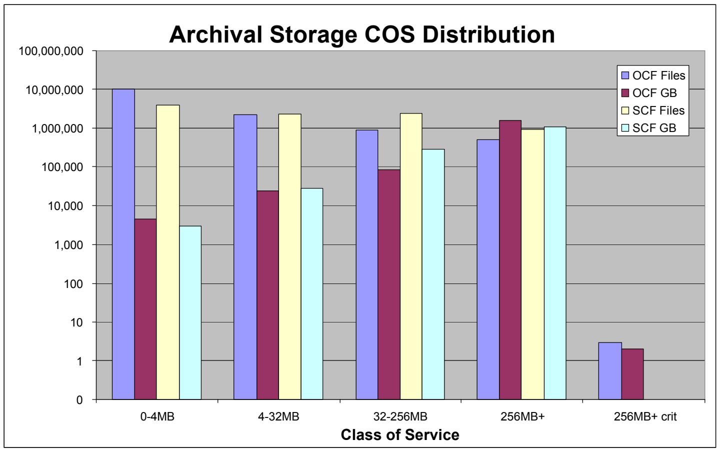


Figure 6: FY06 File and Data Storage by COS

With disk prices dropping and capacity ever-growing, the capacity of the small file disk cache will be increased to be able to store all HPSS small files (under 4MB) online. This strategy has several benefits: it minimizes the problems associated with storing/retrieving small files on tape, and it minimizes stage time for the files which are most-read by users. Aggressive repack will be required to stage millions of existing HPSS small files from tape to the new disk in order to fully realize these benefits. The dual-copy strategy for medium-size files (4-32MB) will remain unchanged as the required disk capacity and cost is prohibitive.

HPSS Movers

Each mover node is connected to the External network (e.g., serial FTP), Jumbo Frame network (e.g., PFTP), and Migration network. Additionally, a mover node is directly connected to disk or tape IO devices, or more typically to the SAN with 1 or more FC HBAs. These connections are shown in ***Figure 4***.

All disk movers are IBM 6Mx machines running AIX. Each disk mover node is designated for one of two groups of Class of Services (COS): COS for <256MB and COS for 256MB+ ("jumbo"). The Jumbo disk mover configuration is optimized for throughput. Each have four FC HBAs directly attached to all 4 ports on the DDN RAID controller. The other disk movers (the COSes with files <256MB) are optimized more for capacity than throughput. Each have 8 FC HBAs connected to the SAN to 16 ports of RAID controllers. DDN RAID controllers achieve about 600MB/s for reads and writes.

The majority of the tape movers are IBM 6Mx machines running AIX. These IBM tape movers each have 4 FC HBAs, each connected to a FC switch with 5 tape drives – a total of 20 tape drives per tape mover machine. The STK 9840 tapes are capable of 15MB/s compressed (10MB/s before compression). The STK 9940B tapes are capable of 45MB/s compressed (30MB/s before compression). LTO3 drives are capable of 120 MB/s (80MB/s before compression while our new Sun Titanium T10000 drives are capable of 180MB/s (120MB/s before compression). Starting in FY06 Linux tape movers were introduced into limited production. Linux tape movers will have a single, dual-port FC4 HBA with each port connected to one tape drive (either LTO-3 or Titanium). We'll continue using the Brocades (for fail-over ability) but the plan is to eventually move to direct-attached tape drives to minimize the cost of FibreChannel infrastructure.

Understanding that the disk and tape mover nodes provide about 130 (jumbo frame) Gigabit Ethernet connections to the two HPSS switches (shown in ***Figure 4***, above the mover node icons), there is network bandwidth oversubscription in the trunks between the HPSS switches and the core network switches (top row of 3 switches in the figure). Currently, there are two 10GigE trunks connecting the HPSS switch on the left (for the disk movers) and one 10GigE trunk for the switch on the right (for the tape movers). Considering everything, this results in approximately a 6:1 network bandwidth oversubscription between the Core switch and the HPSS movers. This trunking bandwidth may be inadequate, particularly for the tape movers which require a sustained data flow to maintain streaming tapes.

For FY07, the number of mover nodes, peripherals and size and connections to the HPSS disk cache will increase to provide specified increases in throughput and to accommodate the retirement of obsolete hardware. Linux tape movers will also be introduced, along with the STK Titanium drives. Since tape drives perform best when streaming, it is important to maintain data throughput to/from the tape drives to keep them spinning. As such, these tape drives are able to adjust their speed (e.g., 2-5 speeds) to best match the source (tape write) or sink (tape read) throughput of the data.

SAN (Storage Area Network)

The SAN uses a combination of 2Gbps and 4Gbps Fibre Channel (FC) to provide connectivity between HPSS movers and tape and disk controllers. The SAN strategy for disk and tape is different. HPSS disk movers for the jumbo (256MB+) file COS are directly attached to disk controllers and hence use no SAN. HPSS disk movers for small through large file COS use a SAN of smaller FC switches (Brocade models 3800 and 4100) configured to aggregate disk controller ports to a fewer number of HBAs at the movers (via FC zoning). If a failure occurs in a mover, disk controller or part of the SAN, the FC switches can be quickly reconfigured to maintain access while suffering only a modest decrease in throughput

The FY06 SL8500-mounted tape drives (LTO-3 and Titanium) are connected via large, Brocade 48K Director-class switches. There are 114 drives in the OCF SL8500s and 124 drives in the SCF SL8500s, with a comparable number of host ports that are all being managed at the Brocade, enabling fail-over and re-shuffling of resources. Silo-mounted tape drives (STK 9840 and 9940B) are connected to small FC switches (Brocade 3200) that aggregate 5 tape drives per mover HBA. This is a cost effective solution to provide adequate bandwidth through the mover to the respective SAN attached tape drives.

HPSS Metadata

The HPSS metadata storage is the critical directory that maps between the file system hierarchy and the storage device. Being such a critical element of HPSS, the metadata has two independent servers and storage. The metadata storage is mirrored and located in two geographically diverse locations at LLNL for both OCF and SCF: in B115 and in B451/3. A robust SAN is built to provide this diverse connectivity, and also to provide connections to tape drives dedicated to backing up the metadata.

Lustre, Inter-Cluster (SWGFS)

The goal of the Site Wide Global File System (SWGFS, aka Cluster-Wide File System in ***Figure 2***, and presently implemented as Lustre) is to provide a high performance file system that incrementally scales in throughput and capacity. The SWGFS is shared by all or many of the machines on the site (facility). The scalability requirement has some nuances that are best explained with an example. When a capability (i.e., largest) machine is added to the facility the global file system requirements (i.e., throughput and capacity) of that machine will be met by adding resources to the existing SWGFS implementation, rather than providing all global file system resources with the capability machine. In fact, the throughput requirement of the SWGFS is to "only" meet the peak throughput requirement of the largest machine on site (and NOT the aggregate peak throughput of all machines on site, which would be larger). It is assumed that in practice any machine's peak throughput is of relatively short duration, and occurs infrequently. It is also our experience that capacity requirements are typically met once the throughput requirements are met, due to the rapidly increasing capacity of disks. Also, the SWGFS must deal with several generations of storage products since the SWGFS grows incrementally over time using the latest storage products at the time purchased – a substantial challenge facing Lustre. The SWGFS must also have one metadata server (versus a federation of metadata servers and associated storage systems) since the largest capability machine is expected to meet peak throughput requirements via access to the entire SWGFS.

In the past the hardware and software for the cluster parallel file systems were tightly integrated with the cluster itself and purchased as part of a package deal. With the SWGFS model, integration of a new system would require an incremental upgrade in the capacity and throughput of the SWGFS. Hence the cost of providing a global file system for a machine via SWGFS would be much less than if the machine had a dedicated global file system. The lower cost of the global file system for a machine tracks the lower cost of capability and capacity computing platforms that are increasingly utilizing open (i.e., non-proprietary) software and hardware solutions, such as InfiniBand for interconnects.

The architectural model for high performance file systems is for all cluster-wide (global) file systems to be replaced by the SWGFS, a single global scalable parallel file system serving all the platforms. The rate of progress toward the complete SWGFS is determined by the stability of the Lustre file system as it is deployed and the effort needed to meet full security requirements. Currently the SWGFS, Lustre, is deployed as a set of independent systems that are mounted on machines needing access (see ***Figure 4*** and ***Figure 5***).

Unclassified Lustre OSS (OCF) 28.8GB/s Total					
System	**Storage Capacity (TB)**	**Gateway Nodes / OSS**	**GigE NICs**	**10GigE NICs**	**External I/O**
lscratcha	320	64	2		9.6GB/s
lscratchb	384	32	2		4.8GB/s
lscratchc	384	32	2		4.8GB/s
GT1	192	64	2		9.6GB/s

Table 4: FY07 capacity and configuration for Lustre systems on OCF

Classified Lustre OSS (SCF) 48.6GB/s Total					
System	**Storage Capacity (TB)**	**Gateway Nodes / OSS**	**GigE NICs**	**10GigE NICs**	**External I/O**
Lscratch1 (was GB1)	400	112	2		16.8GB/s
Lscratch2 (was GB2)	400	112	2		16.8GB/s
Lscratch3	960	80	2		15.0GB/s

Table 5: FY07 capacity and configuration for Lustre systems on SCF

The change to the SWGFS model may also bring changes to the computing usage model. For example users will no longer be required to explicitly move data between the computing platforms and the visualization servers. There will also be no reason to keep multiple copies of large files to have them conveniently available. The architecture of SWGFS and its connectivity to the archive is still in the planning stages.

To complete the integration of the SWGFS into the SCF and OCF environments, some convenient access method is required for usage from other than the major platforms. This is expected to be done with one or more NFS portals. With the introduction of NFSv4 over the next several years, it is planned that clients which have the throughput capability will able to use parallel extensions to the NFSv4 protocol to get higher performance to files within SWGFS.

Determining requirements for throughput and capacity of the SWGFS takes some understanding of the file system usage patterns and delivered I/O performance. The benchmark programs used to verify the file system performance for acceptance purposes is very different from most applications. Because of the less optimized file access patterns, applications typically see I/O rates of roughly 25% of the peak. Allowance was made for this difference in the guidelines used to specify the required I/O bandwidth when a new system was procured.

Visualization

Visualization resources provide an essential and critical part of the high performance computing environment. Furthermore, the visualization resources are often used in real-time by the end-user to post-process computational data from the various computing resources. Visualization resources have evolved in response to technologies, computer architectures, and facility architectures:

- The first generation visualization machines were characterized by SGI machines which provided a multi-processor shared memory machine architecture, very high speed hardware assisted graphics rendering, and high speed local disk IO (~600MB/s per thread). These machines were accessed for visualization by only a very few users at a time, reserving one or several of the machine's frame buffers for RGB ("video") image delivery to the end-user's office by special hardware (e.g., Lightwave Communication RGB switch and extenders). This generation visualization servers set a high bar for user expectations. The disadvantages were: expensive system, graphics engines, and peripherals; difficulty in moving data to machine for visualization processing; maintenance was expensive and required skills that were diminishing as the SGI market presence faded.
- The second generation visualization machines were Linux clusters that addressed issues with limited success. The first visualization cluster, Vivid, did not have a cluster file system. It was more of a proof of concept, focusing more on the demonstration of visualization tools. Then followed Sphere and Gvis, which tried the PVC cluster file system with little success. Next the Lustre file system was used for visualization servers: Sphere shares the MCR Lustre system; PVC the BGL Lustre system; and Gvis has a Lustre file system on the classified side. However the throughput and stability, although improving, are a limiting factor in the visualization cluster overall performance. Meanwhile, the visualization tools on clusters nicely complement the visualization tools and capabilities available on the HPC platforms. And CHAOS, Redhat Linux with LC's modifications and management tools, is being run on the visualization clusters. With visualization clusters users run VisIt real time and get the images delivered to their office terminal via the network using commodity components (vs the Lightwave Communication RGB extenders with the SGI), although with fewer frames per second.
- The current generation of visualization servers will continue to be based on Linux clusters. Improved performance from Lustre is expected. Until Lustre is deployed "globally" within a facility, a visualization server will likely be associated with a platform and that platform's Lustre system (e.g., Gauss, BG/L and the BG/L Lustre system). Additionally, image delivery still does not meet the frame rate of the RGB extenders for the SGI machines. Another change is more use of InfiniBand (IB) as the interconnect. Reliance on IB as the interconnect also brings the expectation for IB switches to provide 10Gigabit Ethernet ports for connectivity to the Gigabit Ethernet system area network. Associated with IB and the IB-Ethernet gateway functionality is the reliance on the success of OpenIB. Visualization cluster deployments recently completed or in progress are: Vertex (16 dual Opteron nodes, InfiniBand interconnect), Klein (10 dual Xeon nodes, Elan4 interconnect), and Gauss (256 Opteron nodes, InfiniBand interconnect) that will follow BG/L to classified and share BG/L's Lustre file system.

Interconnect and System Area Networks

Proprietary interconnect solutions will continue to be a critical part of leading edge computers such as IBM's Purple and the BlueGene/L system. However, large Linux clusters at LLNL have been using solutions that are not proprietary to the machine vendor, such as Myrinet, Quadrics and even Ethernet in a few cases.

The current interconnect strategy acknowledges the continuing role of proprietary solutions for the leading edge platforms (e.g., BG/L, Purple). However, an industry standard solution is desirable for cluster platforms. InfiniBand (IB) has been selected by the tri-Lab community as the open interconnect strategy for clusters. In support of this strategy, LLNL is participating in IB collaborations for developing standards. Additionally, LLNL is promoting the delivery of high performance IB software stack through participation in the OpenFabric Alliance.

Green Data Oasis

The Green Data Oasis (GDO) is an M&IC funded project - $850K in FY06 - driven by several programs to serve data to an international user community from the Green network. The data on the GDO is expected to grow to several Petabytes over a few years, and will be shared with a large international community at very high throughput. Hence Internet access with a large bandwidth capacity shared by many users is desirable.

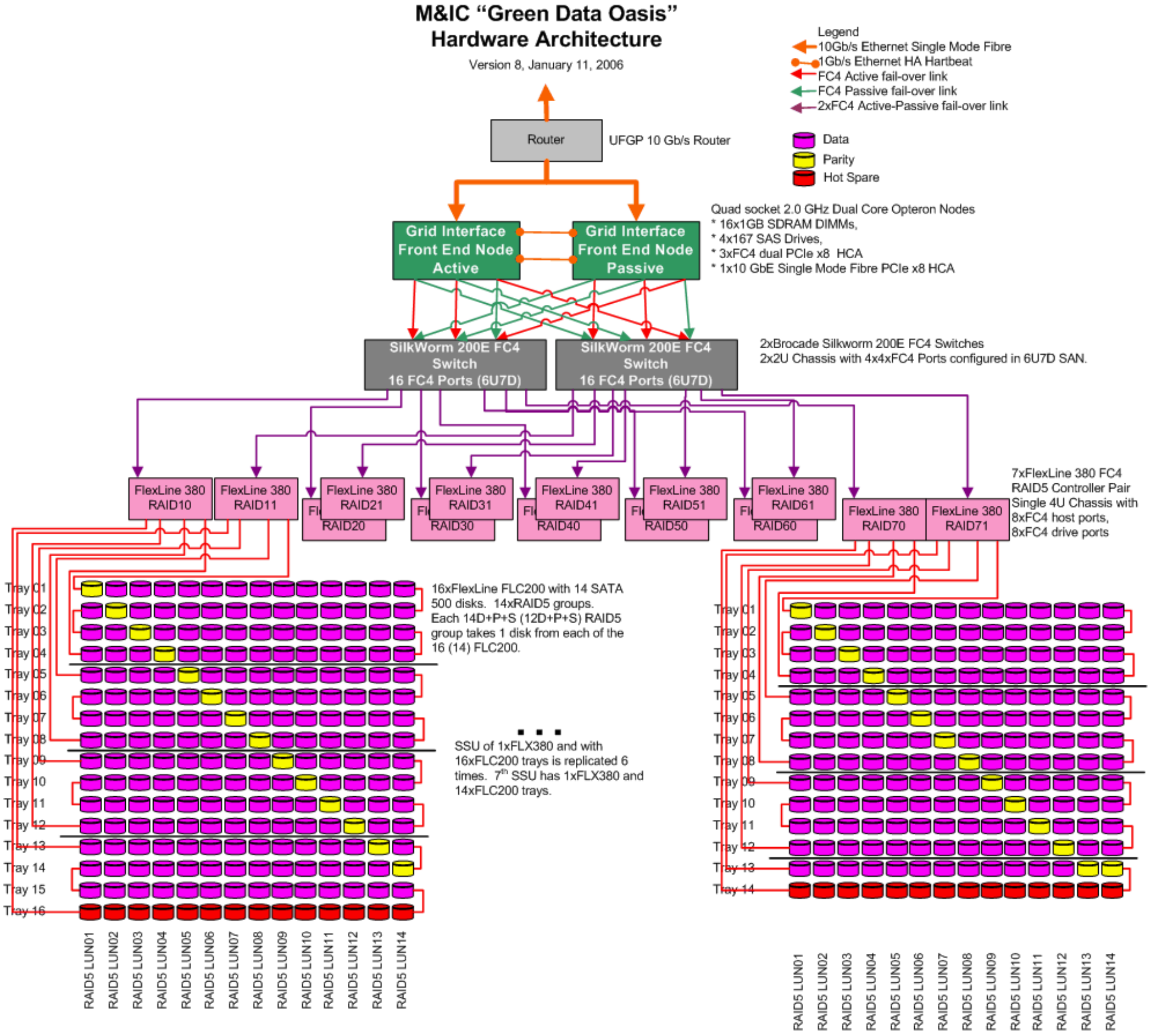


Figure 7: GDO architecture components: ***UFGP Router, GIFEN, FC4 SAN and RAID arrays***

The GDO hardware architecture has four (4) components:

1) University Furnished Government Equipment (UFGP) 10 Gb/s router
2) Grid Interface/Front End Nodes (GIFEN)
3) FC4 SAN Switches and HCA
4) FC4 RAID5 controllers and disk expansion trays

See ***Figure*** *7* for a description of how these four components are connected. Subcontractor shall deliver two (2) GIFEN and seven (7) FlexLine FLX380 Controllers and 110 FlexLine FLC200 disk expansion trays.

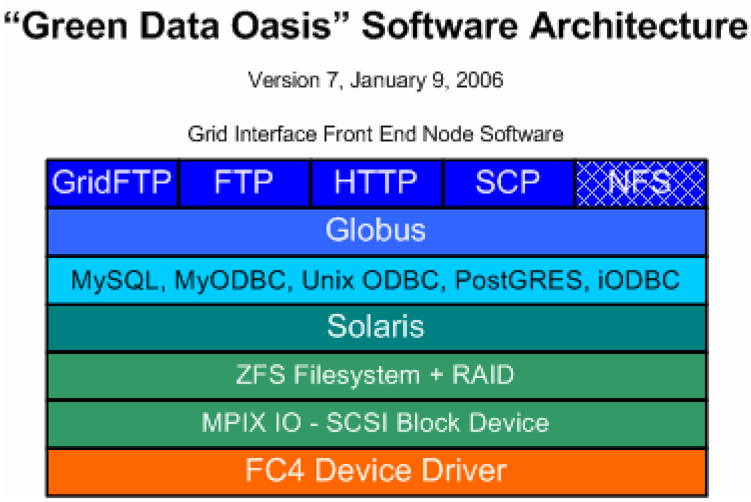

"Green Data Oasis" Software Architecture
Version 7, January 9, 2006
Grid Interface Front End Node Software
GridFTP
FTP
HTTP
SCP
NFS
Globus
MySQL, MyODBC, Unix ODBC, PostGRES, iODBC
Solaris
ZFS Filesystem + RAID
MPIX IO - SCSI Block Device
FC4 Device Driver

Figure 8: GDO Software architecture has 3 components

The GDO software architecture has three (3) components:

1) University Furnished Software (UFS) Globus and GridFTP
2) Solaris including Sun Cluster 3.1 (or later) with HA, FTP, HTTP, SCP and NFSv3 and NFSv4 interfaces and drivers
3) ZFS with HA file system

See ***Figure 8*** for a description of how these three components interface. Sun shall deliver software architecture components (2) and (3). Software components (1), (2), and (3) shall run on the GIFENs.

Status

The Green Data Oasis (GDO) is a large data store on the unrestricted LLNL network intended to facilitate the sharing of scientific data between LLNL projects and external collaborators. The GDO will provide a scalable data repository for roughly 1-2 dozen LLNL projects. The GDO reached LA on August 8, 2006. For the remainder of the calendar year beta users will be making use of the system while the GDO team completes final preparation for GA, which is expected in early January. At that time accounts and disk allocations will be granted through a formal Lab-wide proposal process administered by the M&IC Program Leader with final allocations approved by the LSTO.

Fast network connectivity for the GDO is extremely important to meeting the goals that have been set forth by the LLNL Science and Technology Office. The GDO is on a new segment of the green network, connected as a separate subnet off of the OUTNET edge router. This 1GE network is Phase 1 of the GDO network plan. Once the Lab's connection to ESnet is upgraded to 10GE we will prepare for Phase 2 of the GDO network, which brings it to 10GE by connecting directly to the ESnet router. Phase 3 of the network involves pulling the GDO back onto the main unrestricted network once the lab's internal network has been upgraded to 10GE; this is at least one or two years away.

A critical data path for the GDO is from the LC yellow network to the GDO. Several projects will be generating huge amounts of data on thunder and peloton that they wish to make available for sharing on the GDO. We anticipate from 1-10TB of data per day during the peak data generation period, which could last for weeks. Parallel data movement protocols will be important to minimize the transfer times. Careful attention must be paid to monitoring the network and firewalls to assure overloading does not occur during these peak transfer times.

3) I/O Throughput and Capacity Requirements and Analysis

The IO architecture ideally would be designed so that in totality, and for each system, there would be a perfect balance between 1) network connectivity and bandwidth capacity (e.g., number and aggregate bandwidth of NICs), 2) IO connectivity and bandwidth, and 3) the throughput capacity (e.g., machine's memory, processing and IO performance for all IO applications). This is probably an impractical if not unachievable goal, so compromises must be made. This section will quantify the parameters above for the systems, and lend some discussion and rationale to the decisions reached in the final design and implementation of the I/O architecture of the OCF and SCF.

As a reminder, the following are activities that are sources of requirements (see Section 1 for more discussion):

1) Purple GA on SCF: impact on network, storage, NAS, etc.
2) BG/L to SCF, and/or swings: impact on Lustre, network, storage, etc.
3) Sequoia procurement with assets expected to land in OCF in CY2008, then to SCF.
4) Green Data Oasis.
5) ASC platform plans at other Labs: SNL/NM Red Storm upgrade to100TF, LANL funded for 1PF Roadrunner.
6) Higher speed (10Gbps vs current 2.4Gbps), dual path DisCom WAN deployed in mid-CY2006.

Facility Network

The network requirements will be set in large part by the implementation decisions of the other IO architecture elements: computing resources, storage, NAS, Lustre, and so on. That notwithstanding, some analysis can be done with the information at hand.

The growth in the number of network chassis, 1 GigE ports and 10 GigE ports (see ***Table 1***) is rather significant: about 50% growth in 1 GigE ports and 3500% in growth of 10 GigE ports. This raises several concerns. One is the cost of maintenance (which scales with the number of 65xx chassis), about $370K in FY06 and FY07. Another is the management of these systems, not only for performance and errors but also for configuration control.

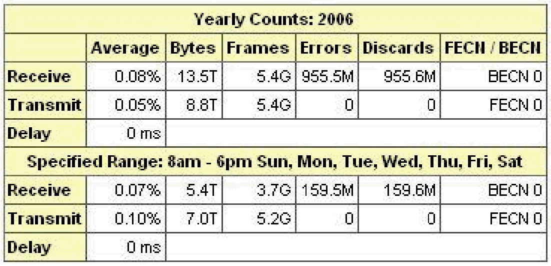
System and NIC Type	B/w (Gbps) per NIC
Ext	0.25
Int (e.g., NFS client)	0.60
1 GigE JF	0.60
10 GigE JF	5.00
Lustre Gwy/Node 1GigE	0.60
Lustre Gwy/Node 10GigE	5.00
1 GigE OSS/OST	0.50
10 GigE OSS/OST	5.00
NFS Server - NetApp	0.40
NFS Server - BlueArc	0.40
NFS Server - Panasas	0.40

Table 6: FY06 per NIC throughput by network and server type

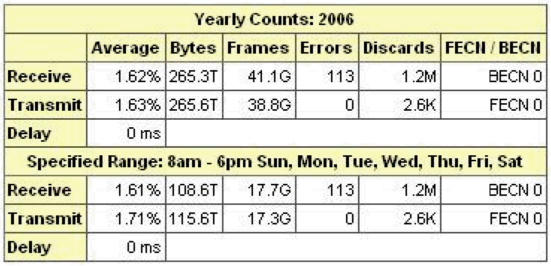
Table 6 (identical to ***Table 2***, but duplicated here for convenience) shows the estimated throughput for the network interface for each type of network and service. These numbers should be re-validated periodically in the IO Testbed using the same hardware, software, and applications that are in production. In FY07, the better NICs, network components and TCP/IP protocol stacks are capable of nearly 100% bandwidth utilization for 1 and 10GigE. However, the applications, IO bus (e.g., PCI and PCI-X versus PCI-express), and legacy NICs typical of hardware and machines purchased and/or designed years ago continue to deliver mediocre throughput as per the table above. We should see dramatic improvements in throughput from the newer machines (e.g., Peloton).

DisCom WAN

The DisCom WAN will grew in FY06 both in bandwidth (from 2.4Gbps to 10Gbps), and connectivity (a link to LANL has been added for additional capacity and redundancy). Also, Gigabit Ethernet IP Taclane encryptors will be used, replacing the ATM UltraFastlane encryptors currently in use. By CY2009 we expect 10 Gigabit Ethernet Taclane encryptors to become available – quite late to make the full 10 Gbps available over both links in a fully redundant architecture. By the start of CY2007 we should have 4Gbps capacity on the DisCom WAN (initial deployment was 2Gbps), using 4 1GigE IP Taclane encryptors at each site.



Yearly Counts: 2006						
	Average	Bytes	Frames	Errors	Discards	FECN / BECN
Receive	0.08%	13.5T	5.4G	955.5M	955.6M	BECN 0
Transmit	0.05%	8.8T	5.4G	0	0	FECN 0
Delay	0 ms					
Specified Range: 8am - 6pm Sun, Mon, Tue, Wed, Thu, Fri, Sat						
Receive	0.07%	5.4T	3.7G	159.5M	159.6M	BECN 0
Transmit	0.10%	7.0T	5.2G	0	0	FECN 0
Delay	0 ms					



Yearly Counts: 2006						
	Average	Bytes	Frames	Errors	Discards	FECN / BECN
Receive	1.62%	265.3T	41.1G	113	1.2M	BECN 0
Transmit	1.63%	265.6T	38.8G	0	2.6K	FECN 0
Delay	0 ms					
Specified Range: 8am - 6pm Sun, Mon, Tue, Wed, Thu, Fri, Sat						
Receive	1.61%	108.6T	17.7G	113	1.2M	BECN 0
Transmit	1.71%	115.6T	17.3G	0	2.6K	FECN 0
Delay	0 ms					

Figure 9: DisCom WAN usage since May 2006 (new DisCom WAN)

DisCom WAN usage is low after the initial testing in April 2006 when the 10Gbps DisCom WAN became operational, with a total of approximately 50TB transmitted and received (see **Figure 9**, and subtracting about 240TB due to testing in May – see figures below).

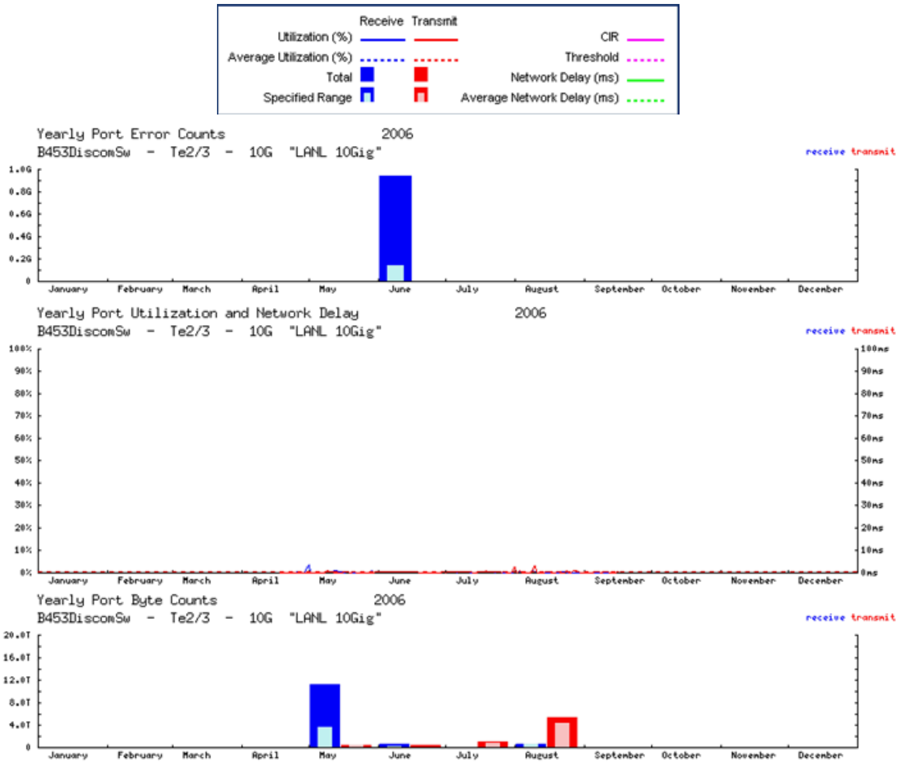


Figure 10: DisCom WAN statistics at LLNL for link to LANL

The network statistics gathered by Statseeker at LLNL for the link to LANL (***Figure 10***) shows a burst of errors in June. This is when the service provider had several problems with their infrastructure resulting in a loss of service for several hours, and also incurring a burst of errors from the network equipment at LLNL. Otherwise, it shows network utilization is quite light.

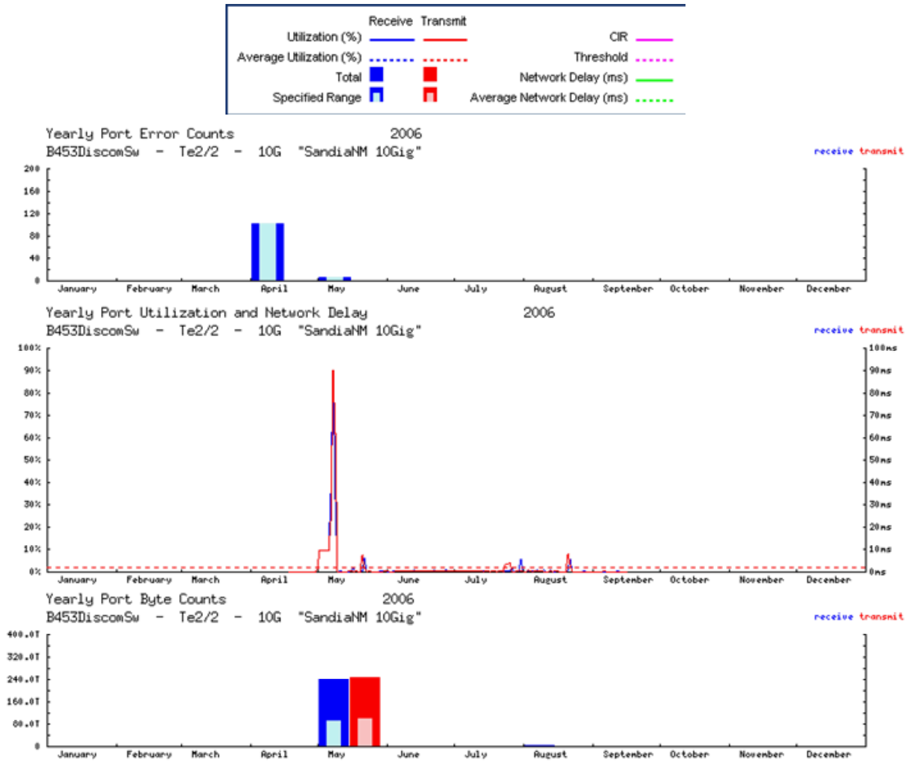


Figure 11: DisCom WAN statistics at LLNL for link to SNL/NM

Here the network statistics gathered by Statseeker at LLNL for the link to SNL/NM (***Figure 11***) shows a few errors in April when the new DisCom WAN was first deployed. Then there is heavy utilization in May during network testing. Otherwise, it shows network utilization is quite light.

In general, as of early FY07 the DisCom WAN usage is extremely light.

Network Tools

In early FY06 LLNL purchased licenses for Statseeker on OCF and SCF. It was deployed on the OCF in late CY06. Deployment of Statseeker on the SCF has stalled due to the security plan. Statseeker runs on an OpenBSD Intel platform, which is a new OS for the SCF hence considerable effort has been brought to bear on getting this new OS validated in the secure computing environment at LLNL. In early FY07 good progress has been made on the security plan, and Statseeker is acquiring some network statistics on the SCF. Already Statseeker has proven to be very valuable in collecting information useful for identifying network performance issues, and applying the appropriate fix or upgrade to resolve those issues. It will be very valuable for SCF too. Usage and performance information on the existing networks allows the network support staff to make informed decisions on upgrading the network and otherwise resolving performance or reliability issues.

Licenses for Opsware were also purchased in early FY06. Opsware is a configuration management tool that was purchased by the institution for configuration management of program and institutional servers and network components. Opsware should be available on the OCF by the end of FY06. The delay was due to the schedule for getting one-time passwords enabled for the Opsware service. Once the network support staff becomes familiar with Opsware on the OCF, it will be similarly deployed on the SCF – probably by mid-CY07.

Network Attached Storage

For the purpose of analysis of NAS requirements we will *exclude systems with global parallel file systems or Lustre file systems*. Doing this, the OCF has 3TF and SCF has 17TF of capability and capacity computing.

Conventional wisdom is to provide 0.0005 Bytes/s per FLOP/s each for productive and defensive I/O. This is a total of 8 Gbps per TF. After excluding the appropriate systems (above), the NAS throughput requirement is 24 Gbps for OCF and 136 Gbps for SCF. ***Table 6*** indicates that the throughput that can be sourced or sinked from the aggregate of all Internal interfaces on all machines (i.e., all NAS clients) is 55 Gbps for OCF and 120 Gbps for SCF, which exceeds or matches reasonably well the requirements above. The estimates in this table also suggest that the current bandwidth of all NAS servers is high by about 50% on OCF and low by 4x on SCF.

	OCF (TB)			SCF (TB)		
	FY04	FY05	FY06	FY04	FY05	FY06
Admin	4.5	7.3	4.9	2.3	4.0	4.0
Home	18.2	18.2	18.2	10.3	10.3	10.3
Project	13.0	13.8	13.8	7.0	7.0	7.0
Scratch	80.0	93.4	89.3	80.0	96.4	96.4

Table 7: NFS storage capacity on OCF and SCF for each of the 4 partitions

Conventional wisdom is to provide at least 20 bytes of globally addressable disk space per FLOP/s or 20GB per TF. Again *excluding systems with Lustre or locally global file systems*, the NAS must cover for 3TF on OCF and 3 TF on SCF. These generate NAS capacity requirements in FY06 of 60 TB for OCF and 60 TB for SCF. The current NAS capacity shown in ***Table** 7* meets the requirements for OCF and SCF.

However, counter to conventional wisdom, the support team has found that the network attached storage capacity and bandwidth scale more by the size of the user community and, after a point, is independent of the number and size of platforms and capacity machines. The rationale is that the network attached storage is used by users for applications and scratch space, and not for data generated by machines which is stored in archive and/or the global file system Lustre. With this rationale, it is not surprising that the capacity and bandwidth of the network attached storage changed little over the past few years, and no significant changes are foreseen for FY07.

Scalable I/O

Generally, the SIO effort addresses file system and customer support issues with I/O rather than determine throughput and capacity requirements. However, this seems a good place to consider the metrics that have been used to determine I/O throughput rate requirements.

Currently, the rule of thumb for GPFS bandwidth is in a TF-to-GB/s ratio with platform performance (i.e., a 100TF-Purple is targeted at ~100 GB/s for its GPFS performance.) The capacity of the file system, in turn, is driven by this bandwidth. This is on account of the number of spinning disks required to attain the performance determines the amount of disk space available. Thus far, this has proven to be sufficient for both bandwidth and capacity.

Another metric might be the amount of real memory for a system that can be written in a fixed amount of time. For example, 50% of real memory written in 3 minutes on 1280 nodes of Purple would require roughly 100 GB/s. Assuming time for opening/closing the file, this would be closer to 3-5 minutes to dump half of memory.

But, one area for consideration might be to determine whether this file system performance is actually required for applications or if optimizations might reduce such requirements. For example, many applications may show 10% of the peak performance as that seen in optimal benchmarking, and few show as much as 50%. It is possible that some do not even scalable as a percentage but rather have closer to a fixed rate due to their design. Considering the cost of throughput and capacity for the parallel file system, it may be better to improve the application design and I/O library performance to get a higher percent of peak file system performance.

Recommendations

Investigate whether bandwidth requirements for future file systems could be reduced by application and I/O library optimization.

Storage

Note in **Figure 12** that while HPSS reads are increasing at a moderate pace, HPSS writes are increasing at a high rate: about 10x every 3 years for OCF and SCF since 2001. Furthermore, 2006 shows a rather steep increase in storage, possibly due to the deployment of BGL and Purple on OCF, and then in mid-FY06 to SCF.

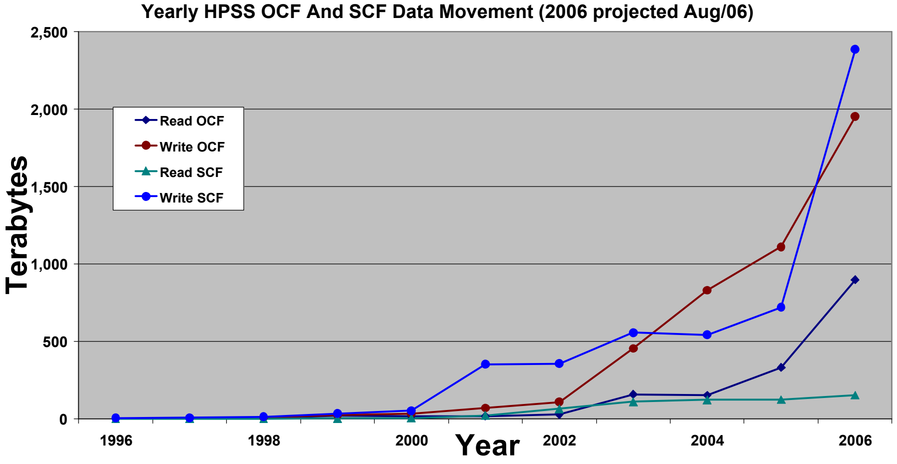
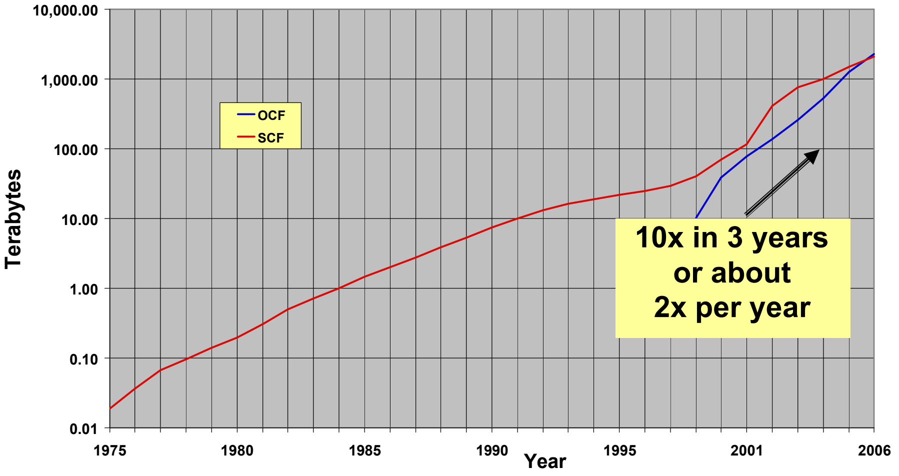


Figure 12: Yearly data read and written to OCF and SCF HPSS

Data storage capacity used for OCF and SCF doubles about every year after 2000 (***Figure 13***, note the vertical grids are spaced every year). This could be tracking the computing capacity in each facility, throughput to storage media, or data movement/management tools. Regardless, an appropriate HPSS storage capacity should be planned and funded.



10x in 3 years
or about
2x per year

Figure 13: Total data in OCF and SCF storage

The tables below (***Table 8*** and ***Table 9***) provide some information on the storage capacity and network connectivity for the HPSS system and movers. This ignores the actual network design, which aggregates the HPSS mover network interfaces to provide a "trunking" bandwidth to the facility's core network (see ***Figure 4***). Using the network throughput noted in ***Table 6*** above, we see that in FY06 the aggregate HPSS throughput from all movers and gateways for clients via the Jumbo Frame network (e.g., parallel FTP, DisCom WAN) is about 80 Gbps. HPSS also provides about 8 Gbps for the clients via the mover's interfaces to the External network (e.g., serial FTP, etc). In FY06 HPSS set and met a "capability" throughput goal of 20 Gbps (2.5 GB/s).

OCF HPSS Storage Characteristics

Parameter	FY04	FY05	FY06
Slot Capacity	5.3 PB	11.2 PB	18.28 PB
Tape Capacity Available	3.9 PB	4.9 PB	12.33 PB
Tape Capacity Used		1.7 PB	3.51 PB
Disk Cache	65 TB	75 TB	100 TB
Archive Nodes	33	33	113
Jumbo Frame GigE NICs	132	132	412
JF Aggregate Peak b/w (Gbps)	12.3 Gbps	79.2 Gbps	247.2 Gbps
Migration Jumbo Frame GigE NICs	132	132	412
Migration JF Peak b/w (Gbps)	12.3 Gbps	79.2 Gbps	247.2 Gbps
External GigE NICs	33	33	113
External Aggregate Peak b/w (Gbps)	8.3 Gbps	8.3 Gbps	28.3 Gbps

Table 8: HPSS Storage Characteristics for OCF

SCF HPSS Storage Characteristics

Parameter	FY04	FY05	FY06
Slot Capacity	4.4 PB	12.2 PB	21.99 PB
Tape Capacity Available	3.4 PB	4.4 PB	11.73 PB
Tape Capacity Used			3.82 PB
Disk Cache	70 TB	75 TB	100 TB
Archive Nodes	36	37	105
Jumbo Frame GigE NICs	144	148	380
JF Aggregate Peak b/w (Gbps)	21.0 Gbps	88.8 Gbps	228.0 Gbps
Migration Jumbo Frame GigE NICs	144	148	380
Migration JF Peak b/w (Gbps)	21.0 Gbps	88.8 Gbps	228.0 Gbps
External GigE NICs	36	37	105
External Aggregate Peak b/w (Gbps)	9.0 Gbps	9.3 Gbps	26.3 Gbps

Table 9: HPSS Storage Characteristics for SCF

Defining HPSS throughput requirements is made difficult because of the highly unpredictable and bursty nature of user-to-archive transfers. The amount of data stored and the bandwidth required is closely tied to the particular applications being run and the personalities and habits of the particular users on the platforms. Add to this the need to be able to handle large data pushes due to file system maintenance or problems, the need for disk-short visualization users to iteratively push and pull data from the archive, and the fact that major platforms have lives on both OCF and SCF, and one can see that a one-size-fits-all strategy for archive requirements is difficult to define. Over the last few years, archive requirements have focused on historical loads, typically describing a doubling in size and bandwidth each year.

An alternative archive requirements strategy, which could be adopted from platforms and Lustre (see Lustre section in Section 2 above), is capability versus capacity. In this context, the capability strategy would be for the goal to meet the requirements for the largest single user/application. The capacity strategy would be to meet the peak aggregate throughput (applying some statistical distribution) of all users/applications. This strategy would then be used to determine the throughput goal to which the network is designed. Specifically the “trunking” between the HPSS aggregating switches and the core network should meet the throughput requirements for HPSS. Since the large HPSS throughput in the tables above is driven less by throughput requirements and more by the design of HPSS (e.g., accommodating COS, striping, storage capacity, etc), the trunking bandwidth provided by the network from HPSS to the facility core network is expected to be less than the bandwidths in the tables above. For example, in FY06 the HPSS movers were connected to the Jumbo Frame network with 80 Gbps throughput, but the throughput goal for HPSS was ¼ that – 20 Gbps (2.5 GB/s). Hence the network trunking from the core to HPSS match the bandwidth requirement of 20Gbps – which it does in FY06 for both OCF and SCF.

Year	Read OCF		Write OCF		Read SCF		Write SCF	
	TB Per Year	Mbps (ave)	TB Per Year	Mbps (ave)	TB Per Year	Mbps (ave)	TB Per Year	Mbps (ave)
1998	3	1	9	2	1	0	10	3
1999	21	5	26	7	3	1	33	8
2000	14	4	33	8	5	1	52	13
2001	17	4	70	18	21	5	352	89
2002	27	7	109	28	63	16	356	90
2003	158	40	454	115	110	28	556	141
2004	150	38	824	209	121	31	542	137
2005	331	84	1110	282	81	21	359	91
2006	789	200	1769	449	174	44	2375	602

Table 10: Trend of data movement by HPSS on OCF and SCF (through 9/30/06)

Over the past 5 years or so the growth of storage capacity has grown on average a factor of over 2 per year (see ***Figure 13***). However, this growth has in fact occurred in jumps – see ***Table 10*** e.g., year 2003 on OCF writes and year 2001 on SCF writes for example. This has likely occurred in response to a new computing resource becoming available (e.g., White on SCF in 2001).

Another point of interest is that on both OCF and SCF the HPSS storage has over 20 Gbps bandwidth to the facility's core network switch but much less bandwidth would be required to move that data on average as indicated in ***Table 10*** e.g., 449Mbps for OCF writes in 2006. This will be discussed further below, referencing network statistics gathered on the OCF in 2006. Similar to all resources in a high performance computing facility, the IO design is heavily influenced by the peak utilization, which was demonstrated – and sustained - in FY06 at 2.5GB/s (20Gbps) for Purple using the same tools available to users. Empirically (in this example), assuming the peak throughput is indeed adequate for users, the peak to which the IO architecture is designed is about 30 times the actual average (e.g., from the table above for writing on SCF in 2006 the ratio is 20Gbps:602Mbps = 33:1 for 2.4PB over a year) throughput – a potentially interesting design parameter which perhaps could be used to correlate between increase in storage capacity and (peak) storage throughput.

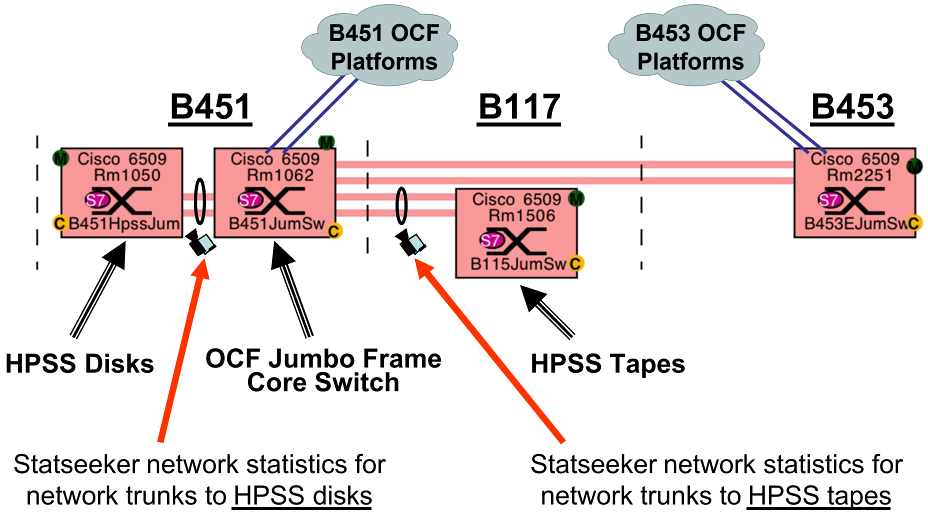


Figure 14: Network traffic statistics gathered by Statseeker

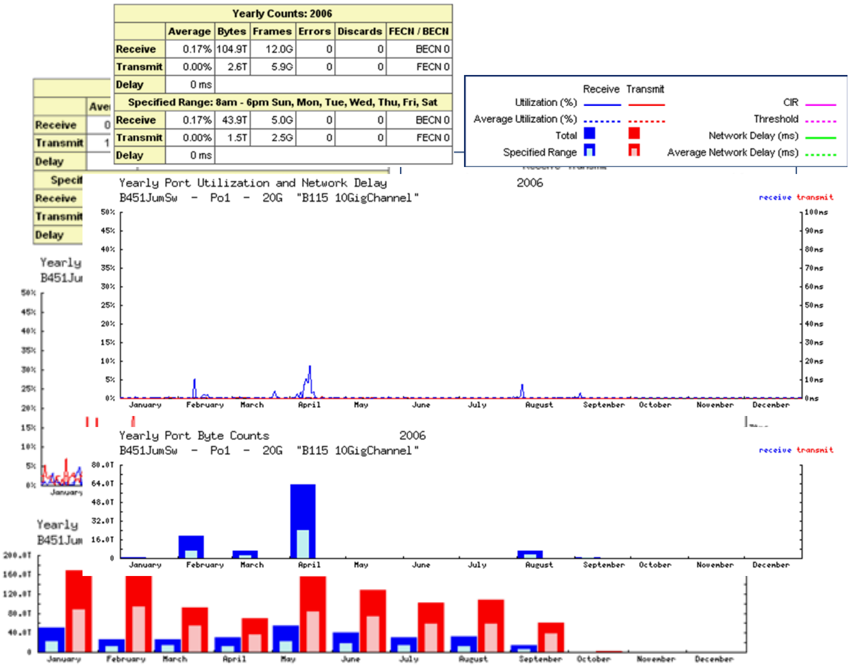
Network Statistics

Early in FY06 LC purchased and deployed on OCF a new network statistics tool, Statseeker. The SCF licenses have been purchased, but deployment is pending approval of the security plan. This tool has been very helpful in understanding the actual utilization of the network, and by extension the utilization of the resources on the network. In this section, information gathered for FY06 on the OCF for the HPSS archival system (see ***Figure 14***) will be shared and discussed. The statistics in the tables below were gathered on the OCF Jumbo Frame core switch on the network trunks to the HPSS disks in B451, and to the HPSS tapes in B115. Recall the Jumbo

Frame network is the primary network used for high performance data movement (e.g., PFTP, NFT, Htar, etc), but does exclude the data to/from HPSS and the External network (the network statistics indicate this is <1TB, very much in the noise).

Figure 15: From B451 aggregate switch to HPSS tape drives in B115 (Jan thru mid-Sept 2006)

The data above (**Figure 15**) is data from the Jumbo Frame core switch to the HPSS tape resources in B115, for January through mid-September 2006. Note virtually all of the data was received (writes from user perspective on the machines). Further, the peak (over a 5 minute interval) is less than 10% of the 20Gbps bandwidth available on the network trunks (i.e., 2 10GigE links).



Yearly Counts: 2006						
	Average	Bytes	Frames	Errors	Discards	FECN / BECN
Receive	0.17%	104.9T	12.0G	0	0	BECN 0
Transmit	0.00%	2.6T	5.9G	0	0	FECN 0
Delay	0 ms					
Specified Range: 8am - 6pm Sun, Mon, Tue, Wed, Thu, Fri, Sat						
Receive	0.17%	43.9T	5.0G	0	0	BECN 0
Transmit	0.00%	1.5T	2.5G	0	0	FECN 0
Delay	0 ms					

Receive Transmit
Utilization (%)
Average Utilization (%)
Total
Specified Range
CIR
Threshold
Network Delay (ms)
Average Network Delay (ms)
Yearly Port Utilization and Network Delay
2006
B451JumSw - Po1 - 20G "B115 10GigChannel"
receive transmit
50% 45% 40% 35% 30% 25% 20% 15% 10% 5% 0%
100ms 90ms 80ms 70ms 60ms 50ms 40ms 30ms 20ms 10ms 0ms
January February March April May June July August September October November December
Yearly Port Byte Counts
2006
B451JumSw - Po1 - 20G "B115 10GigChannel"
receive transmit
80.0T 64.0T 48.0T 32.0T 16.0T 0
January February March April May June July August September October November December
Yearly
B451Jum
200.0T 160.0T 120.0T 80.0T 40.0T 0
January February March April May June July August September October November December

Figure 16: From B451 aggregate switch to HPSS disk in B451 (Jan thru mid-Sept 2006)

The data above (**Figure 16**) is data from the Jumbo Frame core switch to the HPSS disk resources in B451, for January through mid-September 2006. The peak (over a 5 minute interval) is less than 30% of the 20Gbps bandwidth available on the network trunks (i.e., 2 10GigE links), with an average utilization (dashed lines, red line is for transmit – user writing data to HPSS from machines) of about 2.5% (or about 500Mbps = 2.5% of 20Gbps). The average utilization in fact agrees very well with the data received from HPSS (see “Write OCF” for 2006 in ***Table 10***). Note that there is considerable headroom between the demonstrated sustained peak throughput of 2.5GB/s (=20Gb/s) to which the HPSS was designed at the time, and an observed peak of 0.75GB/s (=30% peak of 20Gbps).

Also, the amount of data to HPSS disk and tape recorded by Statseeker is about 20% lower than the amount provided by HPSS. This inconsistency is believed due to two factors: the network statistics did not include in this document the transfers over the External data (although the statistics collected indicated this would be about 1% of the total), and secondly, data was lost in Statseeker early in CY2006 due to inexperience with the tool.

Recommendations
Plan to provide adequate storage capacity in FY07 and beyond to meet the rapid growth of storage. In FY07 the OCF computing environment will not increase substantially, therefore, beyond replacement of aging mover platforms, we do not intend to increase archive bandwidth and will focus our investment on tape media to cover a possible doubling in data capacity. In SCF, we plan to absorb the full production operation of Purple, BG/L and Peloton clusters via a doubling of both bandwidth and archive capacity. During the year we will consider how to change user behavior, tools, and facility design to curb this exponential growth e.g., consider how Lustre (a SWGFS available to every computing resource in the facility) should change user behavior regarding file

archiving, consider a re-compute strategy, re-consider quotas, etc. We also need to begin planning for hosting the initial Sequoia system archive requirements in the OCF.

Lustre, Inter-cluster (Site Wide Global File System)

I/O requirements (i.e., throughput and capacity) of the Site Wide Global File System (SWGFS, aka Cluster-Wide File System in ***Figure 2***, and presently implemented as Lustre) are that of the single, largest machine. The throughput requirement of the SWGFS is to only meet the peak throughput requirement of the largest machine on site (and NOT the aggregate peak throughput of all machines on site). It is assumed that in practice any machine's peak throughput is of relatively short duration, and occurs infrequently. It is also our experience that file system capacity requirements are typically met once the throughput requirements are met, due to the rapidly increasing capacity of disks. Another outcome of this requirements strategy is that the SWGFS must have one metadata server (or a cluster of metadata servers, versus a federation of metadata servers and associated storage systems) since the largest capability machine is expected to meet peak throughput requirements by accessing the entire SWGFS. This strategy was reviewed and analyzed in considerable detail in FY05, with the recommendation that even though likely to be more expensive, several Lustre file systems should remain in production to mitigate the reliability and availability and operational support (e.g., purging) issues experienced to date. The end result is that Lustre file systems will be consolidated over time to no less than two file systems per facility (e.g., OCF, SCF).

Visualization

Visualization has two requirements in FY07 that are presently unresolved. The first requirement is for Lustre file system space for large visualization data sets that would not be purged for the duration of the several month data analysis period. This requirement will be seen in the next section as an issue for Lustre. The second requirement is an improved image delivery system for high end DNT users to replace the specialized RGB infrastructure. This requirement will not be addressed in the FY07 IO Blueprint since it is not yet a high priority issue, but in the next year or two may become one. Furthermore, it is not considered a technical issue but rather one that must deal with security policies involving DNT to ensure the proper data streams can be established between visualization clusters in LC and DNT desktops. For FY07, a pilot to demonstrate a solution may be considered that would also highlight the security policy issues.

4) Issues, Analysis and Recommendations

Meta-issues for FY07 and into FY08

Archive Strategy for the support of Petaflop computing

The anticipated addition in upcoming years of the Sequoia platforms, two enormous compute engines, brings into question the ability of the archive to handle the data volume and bandwidth that these resources will require. Will the archive be able to handle the load given declining budgets and Data Storage Group (DSG) manpower?

Planning Strategy

The LC has placed a strong emphasis on advanced planning; identifying archive requirements based on computational resources as they become evident. This type of planning helps the center to identify opportunities to make early investments in storage and network technologies that will meet these demanding requirements. This proactive approach affords the LC adequate time to pre-deploy large-scale archive hardware and grow the archive capacity and bandwidth gradually, amortizing the cost and work of expanding the archive. The strategy not only includes building early to meet future requirements, but also verifying that the necessary infrastructure (including network) actually meets the specifications through demonstration of DOE L2 milestones (e.g. the FY05 Purple L2 milestone).

Although the LC has always placed an emphasis on advance planning, storage technology is highly dynamic making it extremely challenging to architect and select hardware ideally suited for meeting the archive requirements before the technologies actually exist. In recent years, the archive procurement strategy has become more agile than in the past, enabling a flexible approach to hardware selection. Just-in-time hardware procurements prevent the archive from being caught in the trap of purchasing technologies that will soon become obsolete.

Preparing For Sequoia

The Sequoia RFP due out in late 2007 is scheduled to deliver about a 0.5PF system in 2QCY08 and a much larger system in 1HCY11. The Archive Storage planning strategy must deliver unprecedented storage capacity and bandwidth to meet the requirements of the Sequoia RFP and the usual growth of computing capacity in the facilities. As noted previously, new capability systems reside in OCF for up to a year before they are swung to SCF. Because of this, Sequoia archive preparations heavily involve both SCF and OCF.

Archive hardware

The DSG must architect the archive within several constraints imposed by factors such as cost, technology, manpower, etc. The most critical constraints are discussed below.

HTAR requires SLICs
The Sequoia machines will continue to generate great numbers of relatively small files following the file-per-processor or even file-per-thread models. Critical to the viability of our archives is the HTAR aggregating user (we see 40 million files bundled inside 4,000 HTAR files). Unfortunately, HTAR can not take advantage of the HPSS Parallel Local File Mover (PLFM) concept which allows a mover to directly pull data from Lustre and place it on a tape drive. Because of this fact, we will always need powerful, high-bandwidth nodes to move data from the platform-owned file systems to storage. For Sequoia, the current plan is to field powerful login nodes purchased with the platforms (unlike BG/L, which required stand-alone Storage-Lustre Interface Cluster (SLIC) machines). While the SLIC approach has advantages in providing users access to the data while the primary machine is offline, it comes at a premium, as the SLIC machine must have adequate bandwidth both to the file system(s) and to HPSS.

Direct-to-tape improves scaling of bandwidth and capacity
Keeping our archive on the same scaling curve as the platforms is difficult. One key architectural change that must take place in order to satisfy the bandwidth and capacity requirements of Sequoia at a reasonable price point is to move to a direct-to-tape architecture for HPSS. Growing the HPSS disk cache to a size capable of satisfying tens of gigabytes per second of bandwidth from the platforms is simply not cost-effective; therefore direct-to-tape mechanisms will be required. In order to make this change intelligently, HPSS usage will be analyzed and the Class of Service (COS) strategy will evolve to better match current usage in terms of file sizes. This must be done carefully as single-file transfer rates must be considered as well. To date, users have enjoyed up to 460 MB/sec on a single, very large file. While moving to a direct-to-tape architecture would increase aggregate throughput to the archive, it could also degrade single-file transfer rates and impact the number of users the archive can deliver service to concurrently (each direct-to-tape user gains control of a number of tape drives that cannot be shared like disk).

To enable the direct-to-tape evolution, the DSG has purchased and begun integrating SUN/StorageTek T10000 (T10K) enterprise-class tape drives, which are capable of delivering 120 MB/sec native and 180 MB/sec (330 MB/sec in STK's lab) at 50% compression (the typical compression ratio at LLNL). This tape drive technology serves our current needs very well with the 1GigE Ethernet infrastructure as the tape drive is able to read and write at Jumbo-Ethernet speeds. The T10K drive speed will increase to an even higher-bandwidth solution with multiple drives on a mover node which will be the strategy when the price-point on 10GigE network hardware drops to a more reasonable level. The T10K drives available for purchase as early as November, 2006 will also have 4Gbps FibreChannel interfaces, providing the capability for even higher bandwidths. Linux movers purchased in FY06 by DSG were purchased with 4Gbps FibreChannel Host Bus Adapters to take advantage of this faster interface as it becomes available.

Networks for Archive sub-system
The HPSS infrastructure has been closely matched to the network in years past. When SKY was the primary Capability machine, HPSS was connected via HPGN switches to provide high-speed transfers between the platforms and storage. As machines like ASCI-White began to be built using jumbo Gigabit networks, the HPSS infrastructure evolved, connecting movers via the same jumbo networks. Now, as

the center begins to move towards 10GigE, the HPSS movers will also begin to gradually move to 10GigE. An informal cost analysis was performed in mid-FY06 and the benefit to cost ratio suggested that the move to 10GigE in FY06 wasn't cost effective. As 10GigE prices drop, the DSG plans to move to a 10GigE network infrastructure, beginning first with movers purchased in FY08 in preparation for Sequoia. The move to 10GigE will greatly simplify our strategy for keeping tape drives streaming on quad-NIC movers. As older technology movers are replaced, their connections will be dropped and the new movers will be purchased with 10GigE. Given current multi-year budget forecasts, DSG is anticipating that all HPSS movers will be connected via 10GigE by the end of FY10. One significant difference that this progression will bring about is that HPSS will no longer rely upon a separate, back-end network to handle migration/stage traffic. All HPSS traffic will be born by the 10GigE network.

In the end, the scalability of the back-end HPSS bandwidth and capacity will be ready for Sequoia anticipated requirements, and in concert with the network team, we'll be able to specify a bandwidth requirement between Sequoia and the archive as appropriate. The bandwidth of the SLIC machine, login node bandwidth, and the network uplink between the platform core switches and the HPSS core switch will be the determining factors to overall aggregate bandwidth to HPSS.

Archive Software

The HPSS product continues to provide an unmatched ability as a scalable archival storage resource. For a decade, HPSS has evolved to allow the LC to successfully scale to meet user requirements through the application of additional hardware. At the same time the HPSS internal infrastructure has evolved with changes to transactional and authentication technologies. With critical thought given to hardware selection and integration, HPSS continues to meet the needs of LLNL's most demanding users.

In order to handle the Petaflop requirements of Sequoia, HPSS will need to evolve even more. HPSS Release 7.1 (R7.1), currently in early development, is sharply focused on satisfying these needs. Increasing transaction rates for small files and implementing file aggregation on tape are key components of this release.

Preparing for Contingencies

Because the budget outlook for ASC in coming years is disconcerting, we must plan for the possibility that we will have to limit the amount of data archived from the Sequoia systems purely out of budgetary necessity. Should the cost of ever-growing archive capacity and performance become cost-prohibitive, the DSG will need to implement a quota mechanism to limit data ingest. Such a mechanism would likely not be real-time, but instead be a background mechanism outside of the HPSS product proper. While it would be unfortunate to have to implement such a quota mechanism, we must prepare for the contingency of not being able to afford unlimited archive capacity.

Manpower

The greatest risk to being able to bring archive hardware and software technologies to bear in support of Sequoia is manpower. The DSG archive team has lost over 4 FTEs worth of manpower without replacement, this while deploying and maintaining more and more hardware, software and fielding unprecedented amounts of user data. We no longer are able to proactively deal with problems, but rather are reactive. In the case of fielding new technology for Sequoia, we not only will be reactive, but there may simply not be enough manpower to accomplish the job.

Remaining Challenges

The fact that initial BG/L and Purple production load was handled in OCF, and again after moving to SCF, does not mean that there are no concerns surrounding existing platforms and their file systems. Of immediate concern is the ability of the SCF infrastructure to provide sustained performance to these platforms. Shortly after the systems swung to the SCF, the DSG rapidly deployed LTO3 tape technology and supporting robotics in the SCF. Record-topping days were demonstrated (41 TB in one day alone) with 300+ TB/month for three months with the majority coming from Purple. Sustained levels of performance will require similarly rapid deployment of T10K tape drives in SCF.

BG/L and Purple have shown that users will continue to generate millions of small files. While many will be aggregated using HTAR, a large portion will not. HPSS Release 7.1 will be focused on small file performance (among other scalability issues). While this work is at least 18 months away from being fielded in production, we will be in significant need of these capabilities by the time they are ready. The DSG will also investigate the feasibility of bringing another tape drive species to bear on this problem that is especially well-suited to small files (IBM's TS1120).

Electrical power is another challenge. Computing and processing resources hold-sway in LC floor space and power decisions, not infrastructure. Available power is running out, or has run out in B451 and B453 archive areas (B451 is of greatest concern). The lack of locally available power may well force equipment moves (because of the requirement to site movers near associated devices). This is expensive and is especially taxing on DSG manpower.

How users actually make use of BG/L, Purple, Sequoia and associated file systems raises many unanswered questions including:

- Will the usage model for Sequoia be different from the expected usage model? As computational resources grow, users might find the need to re-design their codes and in so-doing find that they also need to archive either larger amounts of data, or a larger number of small files.
- Users have recently began to consider the archive as a dynamic, random-access storage space, reading more and more of their data back than in the past (likely due to the fact that there is no longer a dedicated Visualization storage space. See ***Tactical Issues for FY07: Lustre, Inter-cluster Issues, Issue E: Lustre storage for vis data is too transient.***) What effect will increased read demands have on overall throughput to the archive? Should dual-copy file-size boundaries be reconsidered to increase reliability? Will users be satisfied with the lower read-rate requirements that have been agreed upon to date, or will they demand faster access to stored data?
- Is Lustre going to be successful as a single, global file system, or will islands of file system storage continue to be the model? Developing an HPSS-to-Lustre interface (and inter-connecting hundreds of HPSS servers to Lustre) will be considerably different depending on which file system model is used.
- Will the possible deployment of a Tri-lab user interface spark increased usage from offsite computing resources? Will the new tri-lab interface provide a higher peak bandwidth to the archive, making higher capacity needs a reality?

These questions are ongoing topics of discussion in the LC. As BG/L and Purple ramp up production in the SCF, the answers to many of these questions will provide the impetus for significant thrust areas for archive and other infrastructure teams in the LC, and refining the strategy for supporting Sequoia.

Tactical Issues for FY07

Storage Issues

This section discusses archival storage issues to be addressed in FY07. Each issue is described and then recommended actions are presented.

Issue A: Need for SLIC or high-bandwidth login nodes

As described previously in ***Archive Strategy for the support of Petaflop computing***, usage of HTAR to aggregate small files from Sequoia-class machines with file-per-process usage models will be critical to the long-term success of the archive (we see 40 million files bundled inside 4,000 HTAR files today). HTAR's sophisticated client-side buffering algorithms (which are at the heart of its ability to move small files at high-performance speeds) cannot take advantage of the HPSS Parallel Local File Mover (PLFM) concept without a complete rewrite. PLFMs allow a mover to pull data directly from Lustre and write it to a tape drive. Because of this, we will always need powerful, high-bandwidth nodes (in the form of login nodes, or stand-alone Storage-Lustre Interface Cluster (SLIC) machines) to move data from the platform-owned file systems to storage.

FY07 Action/Recommendation

For Sequoia, the current architectural plan is to field powerful login nodes purchased with the platforms (unlike BG/L, which required SLIC machines). The DSG must continue to remind system architects to include funding for high-bandwidth login nodes or SLIC-type machines in future procurements. As part of this year's I/O Blueprint effort we have gained concurrence for this approach with system architects.

Issue B: Direct-to-tape architecture

As described previously in ***Archive Strategy for the support of Petaflop computing***, moving the HPSS archive to a direct-to-tape architecture will be instrumental in satisfying Sequoia's bandwidth and capacity to the archive in a cost-effective manner. Determining how to re-architect the current Class of Service strategy will require careful analysis and trending of current archive usage.

FY07 Action/Recommendation

The DSG will finish analysis of current archive usage and propose a Class of Service strategy that addresses the anticipated needs of Sequoia, including optimal file-size boundaries for direct-to-tape.

Issue C: Quotas

As described previously in ***Archive Strategy for the support of Petaflop computing***, budget outlook for ASC in coming years may prevent the archive from growing capacity and performance fast enough to meet user needs. Should this happen, the DSG will need to implement a quota mechanism in order to control data ingest.

FY07 Action/Recommendation

The DSG will gather requirements and develop a preliminary design for a non-real-time quota mechanism that can be implemented independently of HPSS, as a contingency plan for limiting archive capacity growth.

Issue D: Small file performance

As described previously in ***Archive Strategy for the support of Petaflop computing***, despite the existence of HTAR for small file aggregation, there will always be a need for handling large numbers of small files in the archive. As currently implemented, it will be difficult for HPSS to keep pace with the sheer number of small files and create rates that are targeted for future platform needs.

FY07 Action/Recommendation

The DSG will participate in the collaborative development of HPSS R7.1, which is strongly focused on small file transactional performance improvements and aggregating small files on tape.

Lustre, Inter-cluster Issues

This section discusses Lustre and Inter-cluster issues to be addressed in FY07. Each issue is described and then recommended actions are presented.

In FY06 Lustre file systems were consolidated to realize cost savings in the storage hardware. As shown in ***Figure 17*** for OCF, planned FY07 Lustre deployments will have three Lustre file systems cross-mounted for uBGl, MCR, ALC, Sphere, Zeus, Atlas, Prism, and Thunder. An additional Lustre file system will be connected to Thunder. On the SCF there will be three cross-mounted Lustre file systems for BGL, Lilac, Gauss, Rhea, Minos and Hopi. Overall, the major issues with Lustre, prioritized in order, are reliability, availability (priority the same as availability), and performance.

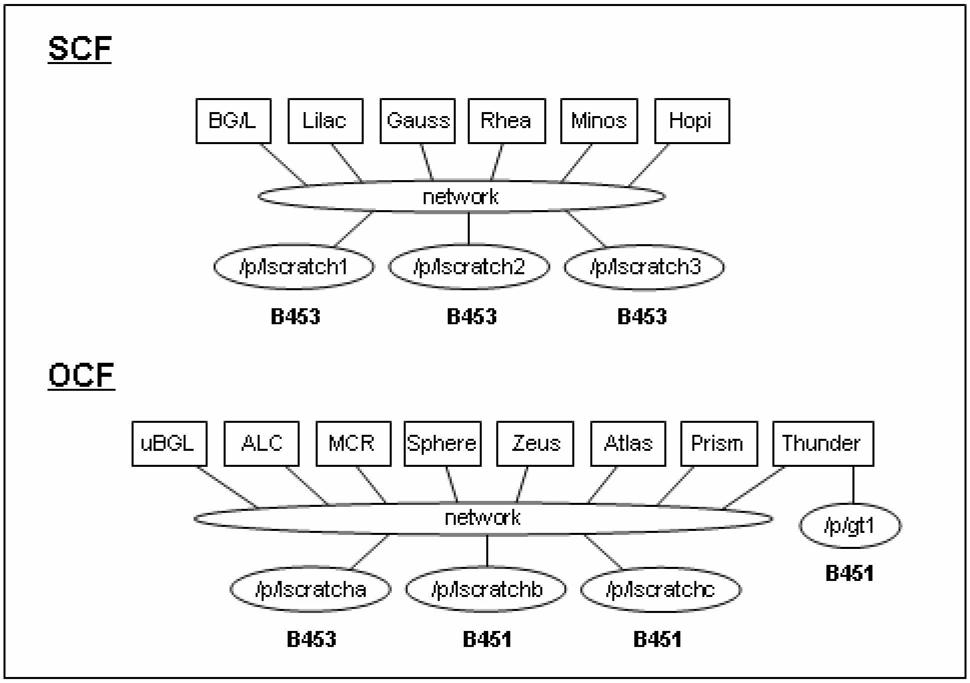


Figure 17: FY07 Lustre Deployments

Issue A: CFS Support
The response from CFS to Lustre bugs has been inadequate. Discussions with CFS identified several potential solutions.

FY07 Action/Recommendation
The new contract with CFS is under negotiation and should address this issue. Some of the suggestions have been: 1) have a CFS engineer on site full time; 2) increase staff to respond to production bugs more quickly; 3) relax turn-around timeframe for bugs in pre-production to give priority to production bugs,; and 4) co-development with LLNL staff for new Lustre features.

Issue B: Operational management of Lustre
Several aspects of supporting and managing the production Lustre systems are difficult, time consuming, or unreasonably impact the operation of Lustre. Examples include: Lustre performance tuning (e.g., 10GigE NICs on Peloton achieve about 9Gbps with the user application netperf, but only get a few Gbps with the Lustre kernel applications – this is considered to be a problem in the Linux kernel), DDN support and management (e.g., local failure procedures), managing OSS/OST failovers, purge policy that currently severely impact users because purges take so long, consolidation of Lustre file systems which often requires moving hardware and reconfiguring Lustre file systems.

FY07 Action/Recommendation
This issue can be traced to lack of resources to develop tools, procedures, and policies. Most recently there has been a spate of hardware failures resulting in a multitude of half to full day file system downtimes that indicates a dire need to analyze and fully understand hardware failure modes and data. We recommend a concerted effort to analyze and understand hardware failure modes, impact and history as well as develop detailed procedures to recover from these failures with minimum impact to users and their data. In addition, we recommend that priority be given to developing and deploying the Lustre software features that would have historically given users the most positive production impact in terms of availability and reliability. This recommendation translates into more Lustre resources.

Issue C: Address complex, long term issues with Lustre
Several issues have been on the hit parade for some time, but are considered risky to put into production (i.e., devastating impact if it doesn't work well) and/or complicated and requiring significant resources and time to address: OST space management (to balance the OSTs, move data from OSTs with little space to OST with more space), adaptive timeouts which would detect higher latency in transactions and

adjust timeouts accordingly (becoming necessary with Lustre cross-mounting), metadata server performance to improve throughput of those transactions, and clustered metadata servers to increase robustness and metadata performance.

FY07 Action/Recommendation
Again more resources for Lustre.

Issue D: Improving Lustre reliability
Making Lustre more reliable, primarily in terms of data integrity and robustness, can be pursued along several avenues: 1) Lustre-level mirroring, wherein the file system handles mirrored data; 2) lower-level mirroring handled by the OSDs; 3) double parity (N+2) RAID to be used by the backend storage; 4) more concerted involvement by CFS (no doubt for a price) in improving Lustre reliability; 5) additional test resources beyond current 50% allotment of ALC (as well as resources when ALC is retired).

FY07 Action/Recommendation
Again more resources for Lustre. In particular for the regular Lustre testing, more of ALC would be beneficial.

Issue E: Lustre storage for vis data is too transient
Several users have generated large data sets with the expectation that they can store it for a long period of time (e.g., months) while doing data analysis. Operational policy is to purge Lustre much more frequently, so these data sets get purged prematurely. Further, it is often difficult for these users to read the large data sets back onto Lustre since the Lustre file system has little available storage space.

FY07 Action/Recommendation
Some suggestions are using quotas to ensure more space is available on the Lustre file systems, and provide Lustre file systems with purge policies more favorable to the visualization user community.

Another possibility would be to provide a Lustre file system for working on large data sets. This file system would reside between the restart dump scratch space and archival storage. Currently, users transfer between scratch space to storage and then back again. The use of a work space would be more efficient. There would continue to be purges on the scratch space, but quotas on the work space.

During the first quarter of FY07, the Pre and Post Processing Environment (PPPE, aka DVS) team had informal meetings with several science run code teams. Though the meetings are still on-going, several conclusions can already be drawn, lead among them that the Lustre environment is failing to meet user expectations, and that this failure has profound impact on both PPPE and storage.

There appear to be several issues that cause the negative user perceptions of Lustre. In no particular order:

- Stat operations (e.g., listing a large directory of files by modification time) are excruciatingly slow.
- Most OCF Lustre file systems operate at 90% or more of capacity, with very little free space available for users.
- Lustre has sub-optimal availability. Jobs are often stalled or quit due to technical problems that hang Lustre.
- Lustre has sub-optimal reliability. Files are often purged, intentionally or unintentionally, so tasks like data analysis that take several months often experience dataset losses.
- Lustre performance for Pre and Post processing suffers due to resource contention between PPPE and compute clusters

Before further discussion, it should be noted that the above are user perceptions, not conclusions based on an objective analysis of the Lustre logs and other data points. Having said that, it doesn't necessarily matter if the points are objectively true; if users believe them to be the case, they will (and in fact do) modify their usage patterns to avoid the "deficiencies" whether real or imagined. And that modification of usage patterns has further impacts on LC services by using systems and services in ways they may not have been designed for.

In the case of Lustre, the original intent was that Lustre would serve as a large, ubiquitous, file store. As such it would be both a place for transient data such as restart dumps as well as for mid-term storage of results that were being analyzed or important data that was being pushed to archive. In this role, Lustre would play a central role in the workflow of most science users, allowing them to share data between compute, PPPE, and storage resources.

The unfortunate reality is that, because science users do not trust Lustre, the HPSS system has subsumed that central role. Lustre is used as a transient file store during compute runs, but results are immediately pushed to HPSS. Data to be analyzed is read from storage back into Lustre, usually multiple times. [Users commented that 2-3 pulls of a dataset was about average, with some making the trip back many more times.] HPSS was not designed to handle this access pattern.

Worse yet, many datasets are "orphaned" in HPSS. In some cases, there isn't sufficient space to pull them back into Lustre for analysis. In others, none of the existing transfer mechanisms are quick enough (or maybe autonomous enough) to avoid auto-logout scripts and other security measures.

On the basis of these conversations, PPPE is looking at expanding the OCF analysis environment by purchasing the equipment necessary to build an additional Lustre file system that would be used for analysis. In concept we have the technology and administrative capabilities to do this. In practice it is cost prohibitive. Using the current designs for Lustre HW, building a file system with sufficient bandwidth to service a viz cluster costs over 3x what the viz cluster costs. [e.g. the Prism viz cluster was approximately $1M and has 128 GigE Lustre connections, building a Lustre file system capable of driving 128 GigE connections costs in excess of $3M.]

Moving forward, the problem seems intractable without a reprioritization of resources. Current FY07 Lustre deployments will significantly increase our storage. But when expressed as a ratio of TB storage to compute teraflops, the change amounts to only less than a 40% increase over historical norms. Given the current usage pattern and its associated impacts on HPSS, this may not be sufficient.

Is there good news? Certainly progress is being made in bringing autonomy to HPSS (and other) transfers via the hopper tool. This in turn may alleviate some of the orphaning issues, but ultimate impact on the science workflows is unclear.

Issue F: Is Lustre still the best choice as a Global File System for LC?

LLNL is committed to providing our users with a High Performance, Scalable, Global file system. Consider “High Performance” to be the I/O rate necessary to support a supercomputer cluster. The ASC Bandwidth to Flops (B to F) ratio suggests a ratio of .001 to .005 (1-5GB/s per TF). “Scalable” suggests that by adding storage and networking capability you should be able to support additional clients. “Global” means that data must be accessible from a large number of independent client systems. Years ago LC selected Lustre as their Global File System. With various issues (e.g., see above) we want to determine if other viable options are available.

FY07 Action/Recommendation

Our recommendation, and in fact course of action, is to complete a File System Bake-off by March 2007. Recently our efforts to provide the Global File System capability have focused on the Lustre file system being developed by Cluster File Systems (CFS) but we periodically attempt to evaluate other products which we believe may satisfy our needs. Because we have a large existing investment in network attached storage hardware, any alternative storage system must be able to use our existing hardware infrastructure in order to be a viable alternative because we don’t have the resources to support two file system products concurrently.

We would like to produce a meaningful side-by-side comparison of production quality file systems which we believe could satisfy our requirements. Our benchmarking activities will focus on performance related issues which can not be evaluated without resorting to a “hands on” approach. Other important topics such as scalability, total cost of ownership, and file system robustness (failover capabilities, tools, etc) will also be discussed as a part of the evaluation process. We will not intentionally induce hardware failures in order to test failover capabilities.

Purple GPFS related Issues

This section discusses Purple GPFS issues to be addressed in FY07. Each issue is described and then recommended actions are presented.

Issue A: Sharing Purple's GPFS File System

On Purple, 64 compute nodes were partitioned off and dedicated as a visualization resource. GPFS has the ability to be mounted on other AIX systems and on Linux systems. If the visualization needs of Purple users could be served by the existing Linux based ASC visualization resources, then those 64 nodes could be returned to a compute partition.

FY07 Action/Recommendation

We have worked with IBM to get a trial license to allow us to test GPFS with Linux clients. The testing will be done at small scale on the OCF network. If successful, the next step would be to develop a security plan to allow for sharing of the file system between systems on the SCF. Also, a plan for a high bandwidth network between Purple and a visualization resource needs to be developed. Both the network and the client licenses would have costs that need to be identified and put into a budget.

More details on the Purple architecture:

The Purple login nodes each have 4 GigE adapters that are un-allocated. Maybe these could be used as a gateway to the Federated network. But, we get about 300MB/s per GigE adapter or about 1.2GB/s per login node for a total of 4.8GB/s. It seems that might be a bit low to make the off cluster use become GPFS viable.

The 120 Purple GPFS server nodes each have 2 copper GigE ports that are un-allocated. If all 240 could be connected into the federated network then the 24GB/s that would provide would be better than the login node option. However, sounds like it would need a lot more network hardware.

Network Attached Storage related Issues

This section discusses NAS issues to be addressed in FY07. Each issue is described and then recommended actions are presented.

Issue A: NFSv4 status

NFSv4 is still early in deployment and not without issues. There have been some compatibility problems found between AIX, Linux, Netapp, and Solaris, specifically in the area of strong authentication and authorization (i.e. Kerberos). In addition, many of the other players in the NAS area like BlueArc and Panasas do not even have a NFSv4 implementation yet.

FY07 Action/Recommendation

Continue to test NFSv4 to get the basic functionality working. In addition, more customer and application requirements will be gathered and tested.

Issue B: NAS vs Lustre for scratch and project directories for smaller capacity machines

Now that Lustre is no longer tied to one compute platform and its visualization cluster, the serial capacity systems are able to mount the same Lustre as the large clusters. In addition, the last non-Linux serial capacity systems have been decommissioned on the OCF; the SCF still has Tempest which is an AIX serial capacity system. With Lustre mounted on both serial capacity clusters and large compute clusters, the need for a NFS scratch and maybe project space is in question.

FY07 Action/Recommendation

The new serial capacity systems Yana and Hopi will be the first that mount Lustre. If this works well for users then no new NFS scratch will be needed in the future. To determine if Lustre works well as a replacement for NFS, scratch we will run some benchmarks of serial I/O. Due to the purge policy, lack of backups and other features like snapshots, the consensus is that Lustre is not a good replacement for project space.

Issue C: Different strategy for network funding

Under the nWBS and the reduced budget, the FY07 network budget was put in FOUS and substantially reduced to fund only projected core network (i.e., the network components interconnecting the various sub-system networks) changes, additions or upgrades.

FY07 Action/Recommendation

Changes or additions to the facilities (OCF and SCF) must now be budgeted to include the required network changes, additions or upgrades. This specifically includes the network supporting storage, NAS, visualization, and the computational machines.

Scalable I/O Issues

This section discusses Scalable I/O issues to be addressed in FY07. Each issue is described and then recommended actions are presented.

Issue A: GPFS and Lustre support

Data integrity, performance (both data and metadata bandwidth), stability, and functionality for GPFS and Lustre parallel file systems require testing and support. Several tools exist currently for this, but the tools' capabilities need to evolve as I/O issues are discovered. Some file system problems are discovered with these tools prior to users encountering them, but some problems are only found and reported by users. In addition to the support for these parallel file systems with regular testing, customer support, covering higher-level I/O library usage, application general I/O design, and file system-specific interaction with customer codes, is vital for making sure the application codes are making good use of the file system.

FY07 Action/Recommendation

Continued efforts on addressing any data integrity problems, in particular developing and enhancing tools to rigorously test the file system, are necessary for both BGL and Peloton Lustre, as well as Purple GPFS. IOR has been used successfully for stress, integrity, and performance testing, and further enhancements to the code have proved important. Metadata testing tools may become necessary for testing and improving metadata performance. In particular, mdtest (MPI-coordinated metadata testing tool developed at LLNL) needs to be enhanced, akin to the effort on IOR last year. Further, continued customer support for user applications on these file systems is imperative: a high-performance parallel file system used awkwardly by applications does not offer much gain.

- Provide necessary enhancements of existing parallel file system testing codes, develop additional tools as needed.
- Continue to support code teams using GPFS and Lustre.

Issue B: I/O library tuning on Lustre

Middle- and higher-level I/O libraries that are used on Lustre may need tuning to work effectively with the underlying file system and with the users codes. These libraries include MPI I/O, Parallel NetCDF, and HDF5. There has not been a concerted effort to study the performance characteristics and tuning possibilities for these libraries on Lustre.

FY07 Action/Recommendation

At this point it is premature to run a study on the higher-level I/O libraries on BGL. While it would have been good to complete such a study and tuning on the open side, the effort on Lustre has rightly been to improve performance, stability, integrity, and functionality of the parallel file system. But while the I/O libraries appear to pose any performance problems currently, studying and tuning of these libraries would likely pay off in performance improvements in the future.

- Monitor I/O library issues to determine when the user's I/O issues are less frequent at the file system level and more often with the library level. Pursue tuning these libraries when it is clear that it offers the highest return on effort.

Issue C: ASC Alliances

We have recently completed two Tri-Lab ASC Level 3 Academic Alliances for I/O research: Northwestern University's (NWU) implementation of distributed caching and University of Michigan's reference implementation of NFSv4. The third Academic Alliance with UC Santa Cruz's research into scalable metadata has been extended (no cost) through fall, 2007.

FY07 Action/Recommendation

We need to maintain our connections that have been developed with these institutions. As a minimum, we must make sure any open employment positions in I/O are advertised to graduating students.

Issue D: POSIX Extensions standardization

There is an effort by the national labs (LLNL, LANL, SNL, ANL), academia (CMU, NWU, UMinnesota), and industry (Panasas, IBM) to improve POSIX performance for parallel I/O by extending I/O calls. Designed more for serial access than parallel, POSIX has shortcomings for parallel I/O that need fixing. While POSIX is not likely to be replaced, extensions to the POSIX standard for parallel I/O are possible. Some of the extensions would relax expensive coherence and metadata semantics, change data movement from streams of bytes to distributed vectors of bytes, allow group locking and file descriptors, inform storage systems of access patterns, and provide improved access control lists (ACLs) for security.

FY07 Action/Recommendation

It seems this is beginning to get a life of its own as more parties get drawn in and vendors in particular begin to show an interest in the possibilities for improving parallel I/O performance. The proposal for these extensions has been developed, and an open group has been established in the process of getting this into the POSIX standard.

- Maintain involvement to drive these extensions into a standardization that vendors provide.

Network Issues

This section discusses network issues to possibly be addressed in FY07. Each issue is described and then described and then a recommended action is presented.

Issue A: 10GigE Taclane encryptors

The DisCom WAN is currently using four 1GigE Taclane encryptors at each site to provide up to 4Gbps throughput over the 10Gbps links. Additional 1GigE Taclanes could be deployed to make more bandwidth available to meet the anticipated utilization (based on utilization statistics gathered at each site). These additional Taclanes add complexity and material and operational costs. Development of a 10GigE Taclane with availability in about 1-2 years will meet our future encryption requirements for the DisCom WAN.

FY07 Action/Recommendation

The development of a 10GigE Taclane encryptor has been on vendor's roadmap for years. There is not yet sufficient demand for this product for vendors to proceed with the development of a product, although considerable progress has been made towards that effort. Our strategy is to continue to make our needs known to the vendors at appropriate conferences and meetings.

Issue B: Large, inexpensive 10GigE switches

The Sequoia RFP will site a Petaflop class machine in 2008 that will require 2000 10GigE switch ports for the global file system (e.g., Lustre).

FY07 Action/Recommendation

Initially LLNL contacted several startups and vendors with a promising roadmap for large 10GigE switches at aggressive prices (i.e., <<$1000 per port). Several startups were targeting early CY07 for initial product availability, and with the slipping of LLNL's Petaflop class machine to 2008 we'll miss that opportunity. So the other Labs have been contacted in the hope that these innovative companies can help meet the broader needs of the tri-Labs.

In the meantime, there is the possibility that additional funding will come to LC to deploy approximately 100TF of capacity computing clusters (Peloton follow-on). For these, we are working with Mark Seager and procurement to develop another RFP, much smaller in scale than the Sequoia Fabric RFP, for a 10Gbps file system area network for the Lustre file system for these capacity computing clusters.

Visualization Issues

This section discusses visualization issues for FY07. Each issue is described and then recommended actions are presented.

Issue A: Network connection outside facility for graphics, data

In addition to requiring significant bandwidth between the Lustre file system and the visualization system, visualization applications must also deliver imagery to user desktops. This can be done in multiple ways: move the data (with FTP and such), move the geometry dataset (e.g., triangles), move the pixels, or by extending the RGB signals from the frame buffers to the desktop (e.g., Lightwave). All options but the last (RGB extenders) result in an additional load on visualization cluster login nodes and the network. There are also security/privacy issues to get network traffic from LC resources to local area networks in DNT's A and B Divisions. As datasets increase in size, the necessary level of interactivity also increases, placing further demands on the network between the user desktop and the cluster.

FY07 Action/Recommendation

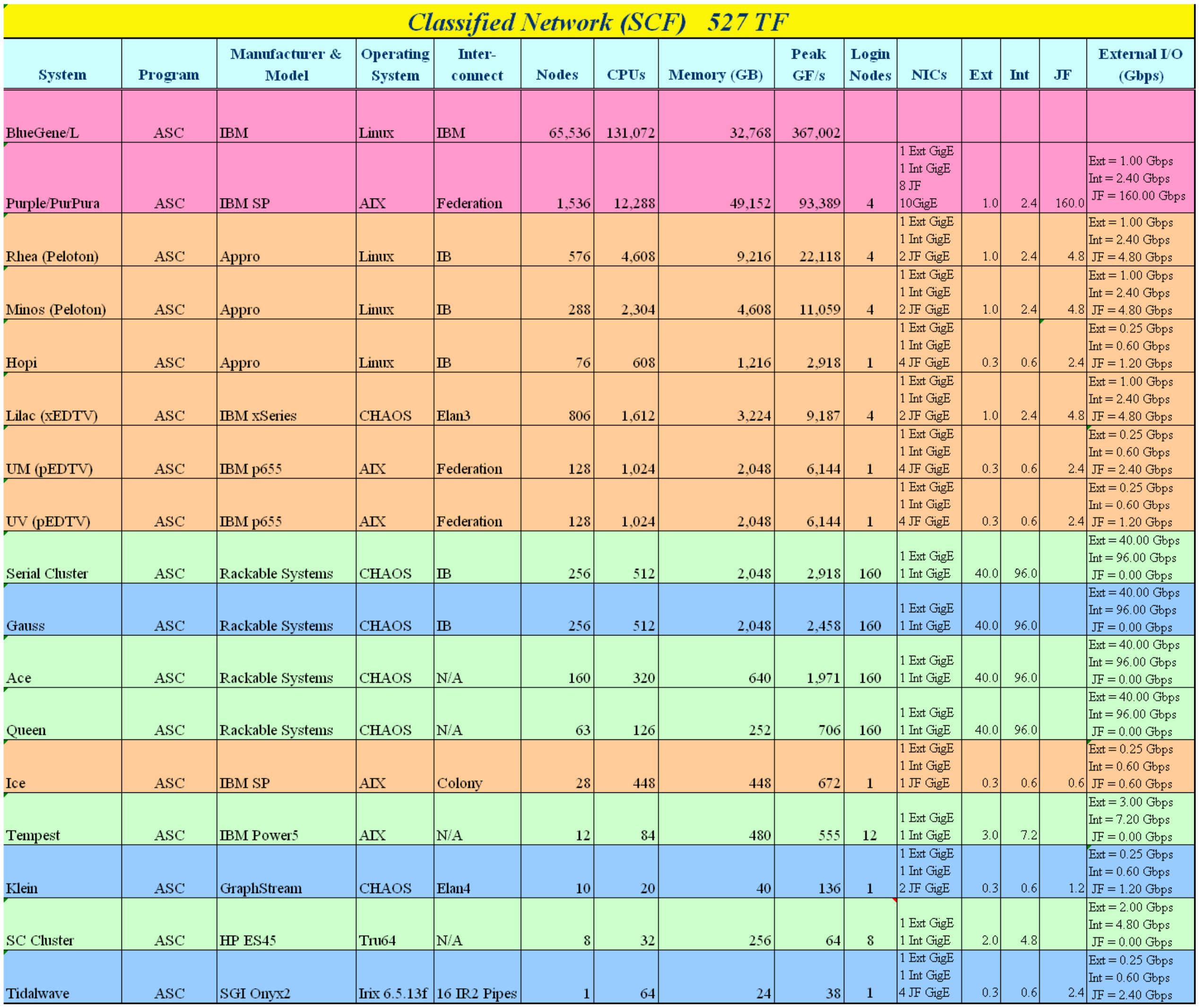
The RGB extenders (from Lightwave) are particularly expensive for visualization clusters (compared to monolithic visualization resources such as SGI), are unreliable, and scale poorly. Moving images or data over the network faces the security/privacy issues, which also tend to impact the performance since the security issues are mitigated by tunneling X-Windows and GLX data through an encrypted SSH session. This remains an unresolved issue which should be addressed soon to open the way for network-based remote visualization solutions in FY08.

There have been demonstrated technical successes for network-based remote visualization. At considerable cost, a few scientists were provided isolated (for security) network connections on the Jumbo Frame network. The network throughput and interactivity allowed the visualization applications to work to the satisfaction of the scientists. This solution does not scale due to cost, and would probably also raise security concerns if too widely deployed. But it does provide a proof of principle.

In FY07, image delivery to the desktop over conventional (e.g., 1 or 10GigE) network components is not technically challenging. However, it does present significant security hurdles since the visualization server is in the LC facility network (e.g., SCF) and the desktop is in the DNT Closed LabNet network. In FY07, the LC visualization leads will begin to address the security challenges for image delivery over Closed LabNet.

Appendix A: FY07 SCF and OCF platform and I/O resources inventory

SCF platform and I/O resources inventory



Classified Network (SCF) 527 TF														
System	Program	Manufacturer & Model	Operating System	Inter-connect	Nodes	CPUs	Memory (GB)	Peak GF/s	Login Nodes	NICs	Ext	Int	JF	External I/O (Gbps)
BlueGene/L	ASC	IBM	Linux	IBM	65,536	131,072	32,768	367,002						
Purple/PurPura	ASC	IBM SP	AIX	Federation	1,536	12,288	49,152	93,389	4	1 Ext GigE 1 Int GigE 8 JF 10GigE	1.0	2.4	160.0	Ext = 1.00 Gbps Int = 2.40 Gbps JF = 160.00 Gbps
Rhea (Peloton)	ASC	Appro	Linux	IB	576	4,608	9,216	22,118	4	1 Ext GigE 1 Int GigE 2 JF GigE	1.0	2.4	4.8	Ext = 1.00 Gbps Int = 2.40 Gbps JF = 4.80 Gbps
Minos (Peloton)	ASC	Appro	Linux	IB	288	2,304	4,608	11,059	4	1 Ext GigE 1 Int GigE 2 JF GigE	1.0	2.4	4.8	Ext = 1.00 Gbps Int = 2.40 Gbps JF = 4.80 Gbps
Hopi	ASC	Appro	Linux	IB	76	608	1,216	2,918	1	1 Ext GigE 1 Int GigE 4 JF GigE	0.3	0.6	2.4	Ext = 0.25 Gbps Int = 0.60 Gbps JF = 1.20 Gbps
Lilac (xEDTV)	ASC	IBM xSeries	CHAOS	Elan3	806	1,612	3,224	9,187	4	1 Ext GigE 1 Int GigE 2 JF GigE	1.0	2.4	4.8	Ext = 1.00 Gbps Int = 2.40 Gbps JF = 4.80 Gbps
UM (pEDTV)	ASC	IBM p655	AIX	Federation	128	1,024	2,048	6,144	1	1 Ext GigE 1 Int GigE 4 JF GigE	0.3	0.6	2.4	Ext = 0.25 Gbps Int = 0.60 Gbps JF = 2.40 Gbps
UV (pEDTV)	ASC	IBM p655	AIX	Federation	128	1,024	2,048	6,144	1	1 Ext GigE 1 Int GigE 4 JF GigE	0.3	0.6	2.4	Ext = 0.25 Gbps Int = 0.60 Gbps JF = 1.20 Gbps
Serial Cluster	ASC	Rackable Systems	CHAOS	IB	256	512	2,048	2,918	160	1 Ext GigE 1 Int GigE	40.0	96.0		Ext = 40.00 Gbps Int = 96.00 Gbps JF = 0.00 Gbps
Gauss	ASC	Rackable Systems	CHAOS	IB	256	512	2,048	2,458	160	1 Ext GigE 1 Int GigE	40.0	96.0		Ext = 40.00 Gbps Int = 96.00 Gbps JF = 0.00 Gbps
Ace	ASC	Rackable Systems	CHAOS	N/A	160	320	640	1,971	160	1 Ext GigE 1 Int GigE	40.0	96.0		Ext = 40.00 Gbps Int = 96.00 Gbps JF = 0.00 Gbps
Queen	ASC	Rackable Systems	CHAOS	N/A	63	126	252	706	160	1 Ext GigE 1 Int GigE	40.0	96.0		Ext = 40.00 Gbps Int = 96.00 Gbps JF = 0.00 Gbps
Ice	ASC	IBM SP	AIX	Colony	28	448	448	672	1	1 Ext GigE 1 Int GigE 1 JF GigE	0.3	0.6	0.6	Ext = 0.25 Gbps Int = 0.60 Gbps JF = 0.60 Gbps
Tempest	ASC	IBM Power5	AIX	N/A	12	84	480	555	12	1 Ext GigE 1 Int GigE	3.0	7.2		Ext = 3.00 Gbps Int = 7.20 Gbps JF = 0.00 Gbps
Klein	ASC	GraphStream	CHAOS	Elan4	10	20	40	136	1	1 Ext GigE 1 Int GigE 2 JF GigE	0.3	0.6	1.2	Ext = 0.25 Gbps Int = 0.60 Gbps JF = 1.20 Gbps
SC Cluster	ASC	HP ES45	Tru64	N/A	8	32	256	64	8	1 Ext GigE 1 Int GigE	2.0	4.8		Ext = 2.00 Gbps Int = 4.80 Gbps JF = 0.00 Gbps
Tidalwave	ASC	SGI Onyx2	Irix 6.5.13f	16 IR2 Pipes	1	64	24	38	1	1 Ext GigE 1 Int GigE 4 JF GigE	0.3	0.6	2.4	Ext = 0.25 Gbps Int = 0.60 Gbps JF = 2.40 Gbps

Table 11: FY07 SCF platform, viz, capacity computing systems

Classified (118 TB)							
Server	Network		Capacity (GB)				Vendor
	Ext	Int	Admin	Home	Project	Scratch	
awing	1	1	358	0	1,461	0	**Netapp**
boba	1	2	0	5,626	0	0	**Netapp**
bwing	1	1	1,978	0	699	0	**Netapp**
chewy	1	2	859	0	1,049	0	**Netapp**
jabba	0	4	0	0		7,331	**Bluearc**
jango	1	2	0	4,679	0	0	**Netapp**
mtd	1	1	0	0	0	0	**Netapp**
solo	1	2	764	0	0	0	**Netapp**
xwing	1	1	0	0	1,858	0	**Netapp**
wampa	20	60	0	0	0	76,283	**Panasas**
yoda	0	1	0	0	0	12,820	**Bluearc**
z4	1	1	0	0	1,921	0	**Netapp**
Totals	**29**	**78**	**3,959**	**10,305**	**6,988**	**96,434**	

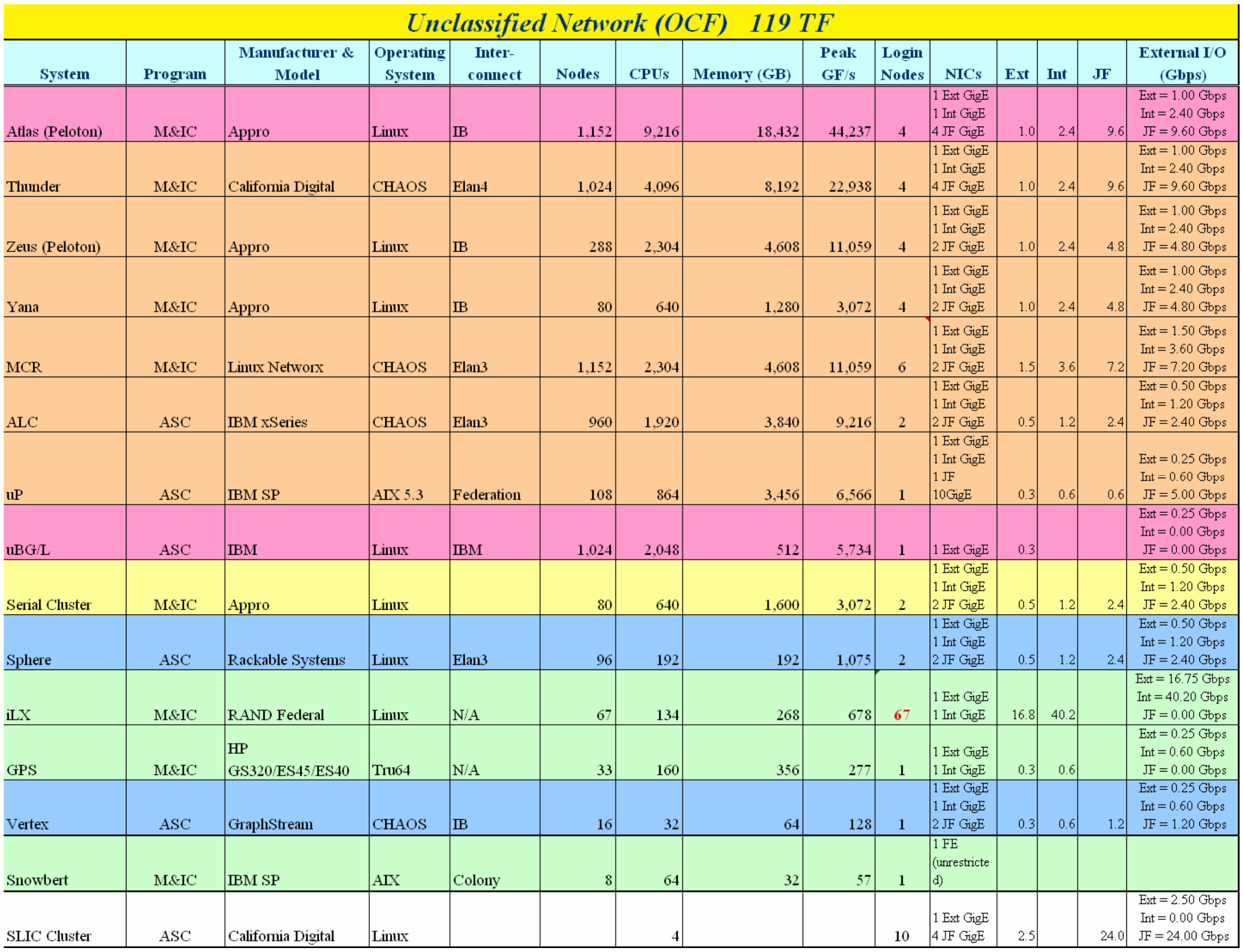
Table 12: FY07 NFS resources[2]

Classified Lustre OSS (SCF) 280 Gbps Total					
System	Storage Capacity (TB)	Gateway Nodes / OSS	GigE NICs	10GigE NICs	External I/O
Lustre GL1	152.0	56	2		56.0 Gbps
Lustre GB1	400.0	112	2		112.0 Gbps
Lustre GB2	400.0	112	2		112.0 Gbps

Table 13: FY06 SCF Lustre resources

[2] All OCF and SCF NFS servers support both TCP and UDP, none are configured for jumbo frame, all run NFSv3, none have NFSv4 enabled, and all have 1GigE interfaces.

OCF platform and IO resources inventory



Unclassified Network (OCF) 119 TF

System	Program	Manufacturer & Model	Operating System	Inter-connect	Nodes	CPUs	Memory (GB)	Peak GF/s	Login Nodes	NICs	Ext	Int	JF	External I/O (Gbps)
Atlas (Peloton)	M&IC	Appro	Linux	IB	1,152	9,216	18,432	44,237	4	1 Ext GigE 1 Int GigE 4 JF GigE	1.0	2.4	9.6	Ext = 1.00 Gbps Int = 2.40 Gbps JF = 9.60 Gbps
Thunder	M&IC	California Digital	CHAOS	Elan4	1,024	4,096	8,192	22,938	4	1 Ext GigE 1 Int GigE 4 JF GigE	1.0	2.4	9.6	Ext = 1.00 Gbps Int = 2.40 Gbps JF = 9.60 Gbps
Zeus (Peloton)	M&IC	Appro	Linux	IB	288	2,304	4,608	11,059	4	1 Ext GigE 1 Int GigE 2 JF GigE	1.0	2.4	4.8	Ext = 1.00 Gbps Int = 2.40 Gbps JF = 4.80 Gbps
Yana	M&IC	Appro	Linux	IB	80	640	1,280	3,072	4	1 Ext GigE 1 Int GigE 2 JF GigE	1.0	2.4	4.8	Ext = 1.00 Gbps Int = 2.40 Gbps JF = 4.80 Gbps
MCR	M&IC	Linux Networx	CHAOS	Elan3	1,152	2,304	4,608	11,059	6	1 Ext GigE 1 Int GigE 2 JF GigE	1.5	3.6	7.2	Ext = 1.50 Gbps Int = 3.60 Gbps JF = 7.20 Gbps
ALC	ASC	IBM xSeries	CHAOS	Elan3	960	1,920	3,840	9,216	2	1 Ext GigE 1 Int GigE 2 JF GigE	0.5	1.2	2.4	Ext = 0.50 Gbps Int = 1.20 Gbps JF = 2.40 Gbps
uP	ASC	IBM SP	AIX 5.3	Federation	108	864	3,456	6,566	1	1 Ext GigE 1 Int GigE 1 JF 10GigE	0.3	0.6	0.6	Ext = 0.25 Gbps Int = 0.60 Gbps JF = 5.00 Gbps
uBG/L	ASC	IBM	Linux	IBM	1,024	2,048	512	5,734	1	1 Ext GigE	0.3			Ext = 0.25 Gbps Int = 0.00 Gbps JF = 0.00 Gbps
Serial Cluster	M&IC	Appro	Linux		80	640	1,600	3,072	2	1 Ext GigE 1 Int GigE 2 JF GigE	0.5	1.2	2.4	Ext = 0.50 Gbps Int = 1.20 Gbps JF = 2.40 Gbps
Sphere	ASC	Rackable Systems	Linux	Elan3	96	192	192	1,075	2	1 Ext GigE 1 Int GigE 2 JF GigE	0.5	1.2	2.4	Ext = 0.50 Gbps Int = 1.20 Gbps JF = 2.40 Gbps
iLX	M&IC	RAND Federal	Linux	N/A	67	134	268	678	67	1 Ext GigE 1 Int GigE	16.8	40.2		Ext = 16.75 Gbps Int = 40.20 Gbps JF = 0.00 Gbps
GPS	M&IC	HP GS320/ES45/ES40	Tru64	N/A	33	160	356	277	1	1 Ext GigE 1 Int GigE	0.3	0.6		Ext = 0.25 Gbps Int = 0.60 Gbps JF = 0.00 Gbps
Vertex	ASC	GraphStream	CHAOS	IB	16	32	64	128	1	1 Ext GigE 1 Int GigE 2 JF GigE	0.3	0.6	1.2	Ext = 0.25 Gbps Int = 0.60 Gbps JF = 1.20 Gbps
Snowbert	M&IC	IBM SP	AIX	Colony	8	64	32	57	1	1 FE (unrestricted)				
SLIC Cluster	ASC	California Digital	Linux			4			10	1 Ext GigE 4 JF GigE	2.5		24.0	Ext = 2.50 Gbps Int = 0.00 Gbps JF = 24.00 Gbps

Table 14: FY07 OCF platform, vis, capacity computing systems

Unclassified (126 TB)								
Server	Network connections (GigE)			Capacity (GB)				Vendor
	Ext	Int	Outside LC Firewall	Admin	Home	Project	Scratch	
bert	1	2	0	0	5,615	0	0	**Netapp**
bigbird	0	4	0	0	0	0	7,342	**Bluearc**
c3	1	1	0	11	0	0	0	**Netapp**
cobalt	1	1	0	0	0	0	0	**Netapp**
count	1	1	0	2,431	0	0	0	**Netapp**
elmo	1	2	1	0	6,163	0	0	**Netapp**
ernie	1	2	0	100	6,409	0	0	**Netapp**
go-ocf	1	1	1	365	0	0	0	**Netapp**
grover	1	2	1	0	0	0	18,166	**Netapp**
mumford	1	2	1	1,976	0	805	0	**Netapp**
rain-lc	1	1	1	0	0	4,418	0	**Netapp**
snuffy	20	60	0		0	0	63,814	**Panasas**
telly	1	2	1	0	0	8,625	0	**Netapp**
Totals	**31**	**81**	**6**	**4,883**	**18,187**	**13,848**	**89,322**	

Table 15: FY07 OCF NFS systems[3]

Unclassified Lustre OSS (OCF) 256 Gbps Total					
System	**Storage Capacity (TB)**	**Gateway Nodes / OSS**	**GigE NICs**	**10GigE NICs**	**External I/O**
Lustre GM1	91.5	64	2		64.0 Gbps
Lustre GM2	91.5	64	2		64.0 Gbps
Lustre GA1	45.0	32	2		32.0 Gbps
Lustre GA2	45.0	32	2		32.0 Gbps
Lustre GT1	192.0	64	2		64.0 Gbps

Table 16: FY06 OCF Lustre systems

[3] See footnotes for OCF NFS table for helpful information.

Appendix B: FY07 computing model

SCF and OCF Computing Model

This section will describe three SCF FY07 computing models. Like last year's models, we have combined capacity and capability computing models into the *LLNL Computing Model* as the influx of powerful Linux compute platforms has sufficiently blurred the distinction between capacity and capability machines. The following assumptions concerning each computing model are critical to the throughput requirements outlined in this document.

The *LLNL Computing Model* is derived from current usage patterns, and user projections. Under this model a mix of large and small data sets are created and stored locally on the machine's disk cache. Local users tend to keep their data on the local disk cache as long as possible and do post-processing, including some software rendering, on the platform. If allowed to complete post-processing on-platform, the data required to be shipped to the visualization servers is reduced. The reduced results will then be stored into the archive from the visualization server. If not allowed to complete visualization on the file system, the data may be repeatedly fetched from the archive for post-processing.

A "typical" example of this model in action (used throughout this document) would involve a user running a week long run generating as much as 40 TB of data on a platform's disk cache. A user needs to get this data off of disk in a "reasonable" amount of time. To get the data off of the machine in one-tenth the time it took to run the problem implies moving 40TB off of the platform and into HPSS in 16.8 hours. This assumes an I/O throughput rate of 661 MB/s. Because most users do not archive all of the data they generate during a run we estimate that a large run may require up to 500 MB/s of bandwidth from the platform to the archive. If it is required to move data off the platform cache before completing post-processing the total data generated will typically be moved to the visualization server for post-processing, necessitating approximately 650 MB/s throughput rate from the platform to the visualization server.

The *LLNL User on a Remote Platform Computing Model* is based upon experience of LLNL users running remotely on the ASC Q machine at LANL. Under this model LLNL users run large simulations on the remote ASC platform often storing interim results and checkpoint saves on the remote site's HPSS archive. Final simulation results may be shipped back to LLNL over the WAN. Visualization of large datasets is performed directly on Q and smaller datasets may be pulled back over the WAN for local visualization and analysis.

The *Tri-Lab ASC User Computing Model* was derived from past Crestone Project runs and discussions with LANL and SNL users. A Tri-Lab user logs into an LLNL ASC machine using his foreign Kerberos credentials. During the problem calculation large data sets are generated and stored on the local platform disk cache. Visualization of data is required as is tertiary storage of selected data sets, and checkpoint saves. LANL users typically take advantage of on-platform visualization greatly reducing the amount of data that needs to be sent over the WAN to their local environments. SNL Tri-Lab users, for the most part, send their datasets back to a Sandia visualization server for rendering locally.

The three computing usage scenarios outlined above are anticipated to be the predominant modes of production operation in FY07, but there will be others. For instance, some users will want to move their full data set immediately to the visualization server or HPSS after computations complete on a platform. From this it can be determined that the paths from the platform to the archive, from the platform to the visualization server and the visualization server to the archive will carry the preponderance of the network traffic.

Looking forward, the system architectural model is likely to change dramatically if the Site Wide Global File System (SWGFS) proves successful. The model would shift from one where each platform has its own cluster-wide file system, to one of a single global scalable parallel file system serving all the platforms. A possible change to the computing usage model as a result of this shift might be that users will no longer be required to explicitly move data between the computing platforms and the visualization servers.